SRAM DESIGN FOR THROUGHPUT ENHANCEMENT

A Thesis Submitted

In Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

by

AYUSH

(Roll No. 2K21/PHDEC/19)

Under the Supervision of

Supervisor

Prof. Poornima Mittal

Professor, ECE, DTU

Co-Supervisor

Prof. Rajesh Rohilla

Professor, ECE, DTU



Department of Electronics and Communication Engineering

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road

Delhi-110042. India

May, 2026

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CANDIDATE'S DECLARATION

I, **Ayush (2K21/PHDEC/19)**, hereby certify that the work which is being presented in the thesis entitled **"SRAM Design for Throughput Enhancement"**, in partial fulfilment of the requirements for the award of the Degree of **Doctor of Philosophy**, submitted in the Department of **Electronics and Communication Engineering**, Delhi Technological University is an authentic record of my own work carried out during the period from **August 2, 2021** to **May 20, 2026** under the supervision of **Prof. Poornima Mittal**, Professor, Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India, and co-supervision of **Prof. Rajesh Rohilla**, Professor, Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India. The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Date: May 20, 2026
Place: Delhi, India

AYUSH
(Roll No. 2K21/PHDEC/19)

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor(s) **Signature of External Examiner**

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE BY THE SUPERVISOR(S)

Certified that **Ayush Dahiya (Roll No. 2K21/PHDEC/19)** has carried out his research work presented in this thesis entitled **"SRAM Design for Throughput Enhancement"**, for the award of **Doctor of Philosophy** from the Department of **Electronics and Communication Engineering**, Delhi Technological University, Delhi, under our supervision. The thesis embodies results of original work, and studies are carried out by the student himself, and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Signature

PROF. POORNIMA MITTAL
Supervisor
Professor
Department of ECE
Delhi Technological University
Delhi, India

Signature

PROF. RAJESH ROHILLA
Co-Supervisor
Professor
Department of ECE
Delhi Technological University
Delhi, India

Date: May 20, 2026

ABSTRACT

The rapid growth of data-intensive computing has increased the performance pressure on memory systems. In conventional processor–memory organizations, a significant fraction of energy and latency is consumed in transferring data between memory and logic units, even when the required operation is simple. Static random-access memory (SRAM) remains a key on-chip memory technology because of its high access speed, compatibility with standard CMOS processes, and widespread use in cache memories, register files, embedded buffers, and local storage blocks. However, SRAM design becomes increasingly difficult under supply-voltage scaling, process-induced variation, bit-line loading, sensing delay, and energy constraints. These challenges become even more important when SRAM is extended from conventional storage to SRAM-based in-memory computing (SRAM-IMC), where the memory array is also used to perform local computation. SRAM-IMC offers a promising route to reduce repeated data movement by allowing selected logic operations to be performed within, or close to, the SRAM array. In this context, throughput enhancement cannot be treated as only a cell-level or macro-level problem. It requires coordinated design of the sensing path, bit-cell, compute macro, and array organization. The present research work therefore focuses on SRAM design for throughput enhancement using a circuit-to-array design approach.

This thesis presents a unified SRAM design framework across four linked technical levels: sense amplifier, SRAM bit-cell, SRAM-IMC macro, and reconfigurable SRAM-IMC array. The sense amplifier is studied first because reliable and fast sensing is essential for both conventional SRAM readout and SRAM-IMC output generation. The SRAM bit-cell is then investigated because storage stability, write ability, half-select immunity, and low-voltage robustness originate at the cell level. The scope of the work is further extended to SRAM-IMC macro-level Boolean logic, where stored operands are used directly in the memory structure to reduce data movement. Finally, a reconfigurable SRAM-IMC array is developed to evaluate whether the proposed memory-compute approach remains useful when operated in both memory and compute modes under array scaling. Thus, the thesis treats SRAM throughput enhancement as a co-design problem involving sensing margin, bit-cell robustness, local Boolean computation, and array-level scalability.

First, two modified decoupled sense amplifier structures, namely the modified decoupled sense amplifier (MDSA) and the modified decoupled sense amplifier with NMOS foot-switch (MDSANF), are proposed for low-voltage differential SRAM sensing. In both topologies, the regenerative latch is decoupled from the bit-lines during the precharge phase and is enabled only after sufficient input differential has developed. This operation reduces direct bit-line disturbance and improves read decision speed. At V_{DD} = 0.8 V and C_{BL} = 50 fF, the MDSA achieves read-delay reductions of 28.6%, 41.79%, 37.74%, and 30.94% compared with MCSA, DTSA, MHSA, and LSA, respectively. The MDSANF provides comparable delay reductions of 26.13%, 39.78%, 35.58%, and 28.55% with respect to the same reference topologies, while

the NMOS foot-switch provides additional leakage and evaluation-window control. Monte Carlo analysis confirms improved variation tolerance under threshold-voltage mismatch, and layout evaluation shows practical column-level area overheads of 0.297 μm^2 for MDSA and 0.5192 μm^2 for MDSANF. The technology-porting study further shows that moving the MDSA from 22 nm FinFET to 32 nm planar CMOS causes only about a 4.5% increase in mean read delay at V_{DD} = 1 V, while reducing the coefficient of variation from 6.12% to 4.96% and narrowing the delay range by approximately 25%. These results demonstrate that the proposed decoupled sensing approach provides a faster and more variation-tolerant sensing foundation for SRAM and SRAM-IMC designs.

Second, SRAM bit-cell structures are investigated for improved stability, write ability, energy efficiency, and robustness under array access conditions. A half-select-free single-ended 8T SRAM bit-cell, referred to as 8TS, is proposed to improve data retention and prevent unintended flipping in selected, half-selected, and unselected cells. At V_{DD} = 0.8 V, the proposed 8TS cell achieves a static noise margin of approximately 0.232 V, a write margin $WM1$ = 0.515 V, and a read power of 0.186 μW, while limiting layout overhead to 4.86% relative to the 7T reference. Monte Carlo and process-corner analyses confirm that the cell maintains stable write margin and delay behavior under threshold-voltage variation, supply fluctuation, and temperature variation. In addition, a differential 8T SRAM cell, referred to as 8TD, is analyzed and compared with the single-ended 8TS cell. At V_{DD} = 0.6 V, the 8TD cell gives a mean SNM of approximately 0.1932 V with a standard deviation of 0.0056 V, whereas the 8TS cell gives a lower mean SNM of approximately 0.1525 V with a higher standard deviation of 0.0152 V. This confirms that the differential cell provides stronger low-voltage stability and tighter mismatch spread. The results indicate that 8TS is suitable for half-select-free and energy-aware storage, while 8TD is suitable for robust differential sensing and low-voltage SRAM-IMC operation.

The scope of the work is further extended to the design and analysis of an SRAM-IMC macro for bitwise logic implementation. The proposed 8TD SRAM-IMC macro connects two 8TD SRAM cells in a column-wise arrangement to store operands A and B, and uses the shared bit-line behavior together with modified sensing to implement NAND/NOR and XOR/XNOR logic. In this approach, Boolean computation is performed inside or near the SRAM structure without transferring the operands to an external logic unit. Functional simulations verify correct operation for all input combinations, $AB = 00, 01, 10$, and 11. For NAND/NOR operation, 6000 Monte Carlo simulations are performed at V_{DD} = 0.8 V, T = 300.15 K, and ΔV_{TH} = 50 mV. The output levels remain clearly separated for all operand combinations, the worst-case mean delay remains close to 0.4 ns, and the coefficient of variation remains below approximately 2.5%. For XOR/XNOR operation, 2000 Monte Carlo simulations give mean delays of 0.55425 ns and 0.55450 ns, respectively, with standard deviations close to 0.0165 ns. The very small difference between the complementary XOR/XNOR mean delays confirms that the logic paths are well balanced under mismatch. Comparative SRAM-IMC cell analysis further shows that the 8TE structure is more suitable for speed-oriented operation, achieving 381.097 ps latency and 2.624 GOPS throughput at V_{DD} = 0.8 V. The 8TC cell provides the lowest energy per bit, but

with higher latency. The proposed 8TD cell provides energy reduction of about 60%–70% at V_{DD} = 0.5–0.7 V, about 30%–40% near nominal supply, and about 15%–25% at higher supply voltages with respect to the 6T, 8TB, and 8TE high-speed baselines. These results clarify the speed–energy trade-off among the evaluated SRAM-IMC cells.

Finally, a dual-mode reconfigurable SRAM-IMC array is proposed for both memory and compute operations. The architecture integrates the proposed 8TD SRAM cell, modified sensing strategy, and Boolean compute path into a common array framework. In memory mode, the array performs conventional SRAM readout, where the selected bit-cell develops a differential BL/BLB response that is amplified by the sense amplifier. In compute mode, the same bit-line and sensing infrastructure is reused to generate NAND, NOR, XOR, and XNOR logic outputs. The proposed array is evaluated from 2×2 to 32×32 across V_{DD} = 0.6–1.2 V. In memory mode, the static noise margin remains nearly invariant with array size, with relative SNM variation within approximately ±0.6 mV, confirming that storage stability is preserved under the evaluated array scaling. In compute mode, NAND-evaluation latency decreases from approximately 493–517 ps at 0.6 V to approximately 380–388 ps at 1.2 V. Over the same supply range, throughput increases from approximately 2.1–2.3 GOPS to approximately 2.75–2.85 GOPS. These results show that speed and throughput scale favorably with supply voltage and remain relatively stable across the evaluated array sizes. However, the average compute energy per bit increases strongly with both array size and supply voltage. The 32×32 array is roughly 10^2 times more energy-consuming than the 2×2 array, and its energy increases by about 25×–30× from 0.6 V to 1.2 V. This establishes that switched bit-line capacitance, interconnect loading, precharge activity, and peripheral capacitance are the dominant array-level scaling costs.

Overall, this thesis establishes SRAM throughput enhancement as a co-design problem involving bit-cell stability, bit-line development, sense-amplifier timing, local Boolean logic, and array-level energy scaling. The MDSA and MDSANF provide the sensing foundation for fast and variation-tolerant differential SRAM readout. The 8TS cell is preferable for energy-aware and half-select-resilient storage. The 8TD cell is suitable for robust differential SRAM-IMC operation. The 8TE structure is suitable for speed- and throughput-oriented SRAM-IMC. The proposed 8TD cell provides a stronger energy-oriented SRAM-IMC option. The reconfigurable SRAM-IMC array demonstrates that the same memory infrastructure can support both conventional read operation and local Boolean computation. At the same time, the array-level results identify compute energy growth under scaling as the main direction for future optimization. Thus, the present research work provides a circuit-to-array SRAM-IMC framework for low-voltage, energy-aware, and throughput-oriented memory-centric computing.

ACKNOWLEDGEMENTS

I take this opportunity to express my sincere gratitude to all those who have supported and encouraged me during the course of my doctoral research. The completion of this thesis has been a long and meaningful academic journey, and it would not have been possible without the guidance, cooperation, and constant support of many individuals.

I am deeply grateful to my supervisor, Prof. Poornima Mittal, Professor, Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India, for her invaluable guidance, continuous encouragement, and generous support throughout my research work. Her academic insight, constructive criticism, and careful supervision have helped me improve the quality and direction of this thesis. I am especially thankful for the time she devoted to reviewing my work, discussing research problems, and guiding me whenever I faced difficulties. Her patience, motivation, and confidence in my abilities have been a constant source of strength during this doctoral journey.

I would also like to express my sincere thanks to my co-supervisor, Prof. Rajesh Rohilla, Professor, Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India, for his valuable suggestions, technical guidance, and consistent support. His thoughtful feedback and critical observations helped me refine several aspects of my research work. I am grateful for his encouragement and for the time and effort he invested in guiding me during different stages of this thesis.

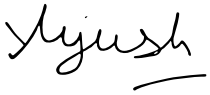
I extend my sincere gratitude to the Head of the Department, the faculty members, and the staff of the Department of Electronics and Communication Engineering, Delhi Technological University, for providing the academic environment and facilities required to carry out this research work. I am also thankful to the DRC chairperson and the respected members of the Departmental Research Committee for their valuable comments, critical assessment, and useful recommendations during the progress of my research. Their suggestions helped improve the clarity, presentation, and technical quality of this thesis.

I am thankful to my seniors, colleagues, and fellow research scholars in the Department of Electronics and Communication Engineering for their support, cooperation, and useful discussions throughout this journey. The exchange of ideas, technical conversations, and research-related interactions with them helped me in several ways during simulations, manuscript preparation, thesis writing, and other academic activities. Their presence made the research environment more engaging and supportive.

I would like to express my heartfelt gratitude to my parents and family members for their unconditional love, patience, and constant encouragement. Their blessings, sacrifices, and support have been the foundation of my strength. They stood by me through every phase of this journey and motivated me to continue working with sincerity and dedication. I am deeply thankful for the values, discipline, and confidence they have given me, which have played an important role in the completion of this thesis.

I also thank my friends and well-wishers for their understanding, encouragement, and moral support during this research journey. Their presence helped me remain motivated during difficult phases and maintain a positive outlook throughout the course of my doctoral work.

To all those who have contributed to this thesis directly or indirectly, I extend my sincere gratitude. Every suggestion, discussion, and word of encouragement has helped me in completing this work and has contributed to my growth as a researcher.



AYUSH
(Roll No. 2K21/PHDEC/19)

CONTENTS

Page No.

Title **i**

Candidate's Declaration **ii**

Certificate by the Supervisor(s) **iii**

Abstract **iv**

Acknowledgements **vii**

List of Tables **xiii**

List of Figures **xv**

CHAPTER 1 Introduction and Literature Review **1–36**

1.1 Overview of the Chapter 1

1.2 Fundamentals of SRAM and SRAM-Based In-Memory Computing 2

1.2.1 SRAM Array Organization and Conventional Memory Access 2

1.2.2 SRAM Bit-Cell Operation and Stability Metrics 2

1.2.3 Sense-Amplifier Operation in SRAM Readout 5

1.2.4 SRAM-IMC Macro-Level Logic Operation 6

1.2.5 Array-Level SRAM-IMC Operation and Reconfigurability 7

1.2.6 SRAM-to-IMC Design Hierarchy 8

1.3 Background and Motivation 9

1.4 Literature Review 11

1.4.1 Sense Amplifier Architectures 12

1.4.2 Key Observations from Prior Sense Amplifier Designs 15

1.4.3 Advances in SRAM Bitcell Architectures 16

1.4.4 Key Observations from Prior SRAM Bitcell Designs 21

1.4.5 Development of SRAM-IMC Macro Architectures 22

1.4.6 Key Observations from Prior SRAM-IMC Architectures 25

1.4.7 Array-Level Organization and Evolution of SRAM-IMC Architectures 26

1.4.8 Key Observations from Prior SRAM-IMC Array Architectures 29

1.5 Research Gaps 30

1.6 Problem Statement 32

1.7 Research Objectives 32

1.8 Objective-Wise Methodology 33

1.9 Thesis Organization 34

CHAPTER 2 A Process Variation Tolerant Modified Sense Amplifier for Differential SRAM 37–57

2.1 Overview of the Chapter 37

2.2 Introduction 38

2.3 Proposed Variation Tolerant Modified Decoupled Sense Amplifier 39

2.3.1 Impact of Sizing of Decoupling PMOS on SA Read Performance 39

2.3.2 Parameter Variation Tolerance of Proposed SA 42

2.4 Performance Comparison of MDSA/MDSANF with Existing SA Architectures 45

2.4.1 Classification of SRAM Sense Amplifiers 46

2.4.2 Read Delay 49

2.4.3 Power Consumption and Leakage Power 50

2.4.4 Area Considerations 52

2.4.5 Variation–Tolerance Implications of Porting from 22 nm to 32 nm 54

2.5 Summary of Important Results 56

CHAPTER 3 Proposed 8T SRAM Cell design for Improved Performance 58–86

3.1 Overview of the Chapter 58

3.2 Introduction 59

3.3 Proposed Single-Ended Process-induced Variation Resistant 8TS SRAM Cell 60
3.3.1 Stability Analysis of the 8TS SRAM Cell 62
3.3.2 Write Margin Analysis of 8TS SRAM Cell 64
3.3.3 Power Consumption and Leakage Analysis of 8TS SRAM Cell 68
3.3.4 Read and Write Delay Analysis of 8TS SRAM Cell 70
3.3.5 Half-select Error Resilience of 8TS SRAM Cell 71
3.4 Comparative Analysis of 8TS SRAM Cell with Other Existing SRAM Architectures 73
3.5 Proposed Differential 8T SRAM Cell Architecture 79
3.6 SRAM Stability Margin comparison with single-ended 8TS SRAM 80
3.7 Delay and Power Analysis for Read and Write operations 82
3.8 Monte Carlo analysis for Threshold Voltage Variations 83
3.9 Summary of Important Results 85

CHAPTER 4 8TD SRAM-IMC for Bitwise Logic Implementation 87–106
4.1 Overview of the Chapter 87
4.2 Introduction 88
4.3 In-Memory Computing using Standard SRAM Bit cells 89
4.4 In-Memory Computations using the proposed 8TD SRAM cell and MDSANF 93
4.5 Performance Comparison of In-Memory SRAM Cell Architectures 98
4.6 Summary of Important Results 105

CHAPTER 5 Reconfigurable SRAM-IMC Array for Throughput Enhancement 107–128
5.1 Overview of the Chapter 107
5.2 Introduction 108
5.3 Proposed Reconfigurable SRAM-IMC Array Architecture 110
5.4 Memory-Mode Analysis Under Array Scaling 112
5.4.1 Static Noise Margin 113

5.4.2 Write-Margin Deviation 114
5.4.3 Read Delay 115
5.4.4 Read Power 116
5.4.5 Interpretation of Memory-Mode Results 117
5.5 Compute-Mode Analysis Under Array Scaling 118
5.5.1 Compute-Mode Latency 119
5.5.2 Average Compute Energy per Bit 120
5.5.3 Compute Throughput 122
5.5.4 Interpretation of Compute-Mode Results 123
5.6 Array-Level Scalability and Design Trade-Offs 124
5.7 Summary of Important Results 127

CHAPTER 6 Conclusion, Future Scope and Social Impact 129–135
6.1 Conclusion 129
6.2 Future Scope 132
6.3 Social Impact 134

References 136

List of Publications 155

List of Tables

S. No.	Table Caption	Page No.
Table 1.1	Representative sense amplifier architectures and key trade-offs in design.	12
Table 1.2	Evolution of SRAM cell architectures from conventional 6T designs to multi-transistor cells.	16
Table 1.3	Representative SRAM Array-Level Architectures and Optimization Techniques	20
Table 1.4	Overview of SRAM-Based In-Memory Computing Architectures at the Bit-Cell and Macro Levels	24
Table 1.5	SRAM-IMC Array Designs Focusing on Architecture, Sensing, and Peripheral Overheads	27
Table 2.1	Performance summary of conventional SA designs compared with MDSA and MDSANF at 0.8 V supply voltage	53
Table 2.2	Monte Carlo Read Delay statistics for MDSA (1000 samples, TT corner, V_{DD} = 1 V)	55
Table 3.1	Coefficient of Variance (CoV) Analysis for Write Margin (WM) across Process Corners and Supply Voltages.	67
Table 3.2	Coefficient of Variation (CoV) for Read and Write Delays at Different Temperatures	71
Table 3.3	Area overhead comparison of various SRAM bitcells relative to the 7T cell.	79
Table 4.1	Latency statistics of SRAM-IMC cells under device mismatch at V_{DD} = 0.8 V and ΔV_{TH} = 50 mV.	104
Table 4.2	Representative SRAM-IMC performance at V_{DD} = 0.8 V.	104
Table 5.1	Summary of memory-mode scaling trends for the proposed reconfigurable SRAM-IMC array.	118

Table 5.2 Summary of compute-mode scaling trends for the proposed reconfigurable SRAM-IMC array. 124

Table 5.3 Overall array-level scaling behavior of the proposed reconfigurable SRAM-IMC architecture. 126

List of Figures

S. No.	Figure Caption	Page No.
Fig. 1.1	Basic organization of an SRAM array showing the bitcell matrix, wordline selection through the address decoder, differential bitline pairs, column drivers, precharge circuit, and sense-amplifier readout path.	3
Fig. 1.2	Conventional 6T SRAM bit-cell showing the storage nodes and differential bit-line access.	4
Fig. 1.3	Typical SRAM read sequence showing precharge, word-line activation, bit-line differential development, sense-amplifier enable, and full-swing output regeneration.	5
Fig. 1.4	Memory bottleneck dilemma and architectural alternatives: (a) conventional von Neumann organization, (b) near-memory computing, (c) in-memory computing.	9
Fig. 2.1	Circuit schematics for (a) MDSA and (b) MDSANF, along with the transient response showing timing control signals at V_{DD} = 0.8 V and C_{BL} = 50 fF for (c) MDSA and (d) MDSANF	40
Fig. 2.2	Investigating the impact of PMOS devices M6 and M7 width on MDSA and MDSANF performance: (a) Delay, (b) Power consumption, and (c) Power-delay-product for MDSA and (d) Delay, (e) Power consumption, and (f) Power-delay-product of MDSANF examined for different supply voltages	41
Fig. 2.3	Transient response (10k Monte Carlo simulation) of (a) MDSA, and (b) MDSANF and calculated delay distribution at C_{BL} = 50fF for near threshold operation of (c) MDSA and (d) MDSANF (Inset: delay distribution at above threshold operation)	43
Fig. 2.4	Delay variation as a function of (a) Bitline Capacitance and (b) Process corners for MDSA/MDSANF at VDD = 0.8 V	44

Fig. 2.5	Normalized Delay as a function of differential bit line voltage input for MDSA and MDSANF at (a) near threshold and (b) above threshold supply voltages	45
Fig. 2.6	Circuit schematics for Voltage SAs (a) Conventional latch-type SA, (b) Current-controlled latch-type SA, (c) Decoupled latch-type SA, and (d) Double tail SA.	46
Fig. 2.7	Representative current-mode and hybrid SRAM sense amplifier architectures: (a) clamped bit-line sense amplifier, (b) hybrid current sense amplifier, (c) modified clamped sense amplifier, and (d) modified hybrid sense amplifier.	48
Fig. 2.8	Simulation test-bench used for comparison of sense amplifiers.	49
Fig. 2.9	Delay comparison of all sense amplifiers as a function of the supply voltage at C_{BL} = 50 fF.	50
Fig. 2.10	(a) Power Consumption and (b) Leakage power comparison of all SAs as a function of the supply voltage (C_{BL} = 50 fF)	51
Fig. 2.11	Power-delay-product comparison of all SAs as a function of the supply voltage (C_{BL} = 50 fF)	52
Fig. 2.12	Layout design for area estimation of (a) MDSA and (b) MDSANF	53
Fig. 2.13	Supply-voltage dependence of (a) read delay and (b) dynamic power for the MDSA implemented in 45, 32, and 22 nm CMOS.	55
Fig. 2.14	Power–delay product (PDP) of the MDSA versus V_{DD} for 45, 32, and 22 nm nodes. The crossover at 0.7 V marks 32 nm as the optimal speed–energy operating point.	56
Fig. 3.1	Proposed single-ended, single port 8T SRAM bitcell (A) 8TS circuit schematic, (B) layout of the 8TS bit-cell; Equivalent circuit for (C) hold mode, (D) read mode, and (E) write mode; (F) transient simulation with control signals for all the operations of 8TS Cell.	61

Fig. 3.2 Static noise margin butterfly curve of the 8TS SRAM cell for (A) SNM calculation, (B) butterfly curve for Monte Carlo (2000-point) analysis at V_{DD} = 0.8 V, (C) SNM as a function of supply voltage, and (D) SNM as a function of temperature (Inset: distribution of SNM at T = 27°C for V_{DD} = 0.8 V) 63

Fig. 3.3 Write Margin's calculation for the proposed 8TS SRAM bit-cell (A) WM-0, (B) WM-1, and (C) the impact of supply voltage scaling. 65

Fig. 3.4 Monte Carlo simulations of the proposed 8TS cell across process corners, evaluating write margin (WM) under threshold voltage V_{TH} variations of ±30 mV and supply voltage fluctuations ΔV_{DD} of 25% around the nominal 0.8 V. 66

Fig. 3.5 Power consumption of the 8TS SRAM cell as a function of supply voltage: (A) read and write power variation, (B) leakage power as a function of bitline capacitance, and (C) temperature dependence of write power and (D) leakage power at multiple V_{DD}. 69

Fig. 3.6 Temperature-dependent delay analysis of the proposed 8TS SRAM bit-cell for read and write operations at V_{DD} = 0.8 V. (A) Read delay, (B) Write-1 delay, and (C) Write-0 delay histograms, highlighting the impact of temperature variations on SRAM cell performance. 70

Fig. 3.7 Illustration of half-select resilience in the proposed 8TS SRAM bit-cell. (A) A 2×2 array of 8TS SRAM cells, highlighting the selected, half-selected, and unselected cells. (B) Transient response of control signals and storage node voltages for selected, half-selected, and unselected cells, demonstrating the robustness of the proposed design. 72

Fig. 3.8 Circuit schematics of existing SRAM bit-cells with varied transistor count (A)7T, (B) 8Ta, (C) 8Tb, (D)8Tc, (E) 9T, (F) 10T, (G) 11T, and (H) 12T. 73

Fig. 3.9 Performance comparison of the proposed 8TS SRAM bit-cell with existing 8T SRAM variants: (A) Read/Write delay, (B) Power consumption, (C) Normalized Write Margin (WM) and (D) Static Noise Margin (SNM). 76

Fig. 3.10	Comparative analysis of the proposed 8TS SRAM bit-cell compared to existing SRAM bit-cells with varying transistor counts: (A) Read/Write delay, (B) Power consumption,(C) Normalized Write Margin (WM) and (D) Static Noise Margin (SNM).	77
Fig. 3.11	Layout representations of different SRAM bit-cells (7T–12T) used for comparative analysis, illustrating transistor arrangement and routing complexity: (A) 7T, (B) 8Ta, (C) 8Tb, (D) 8Tc, (E) 9T, (F) 10T, (G) 11T, and (H) 12T. Note: Layouts are not optimized for array-level implementation.	78
Fig. 3.12	Circuit diagram of the proposed differential 8TD SRAM cell	80
Fig. 3.13	Butterfly Curve for calculation of static noise margin (SNM) (a) differential 8TD SRAM cell and (b) single-ended 8TS SRAM cell. (c) Impact of voltage scaling on (c) SNM and (d) Write margin (WM) of 8TD and 8TS cells.	81
Fig. 3.14	Timing analysis of the differential 8TD and single ended 8TS SRAM cells. (a) Read Delay comparison and Write Delay for (b) 8TD and (c) 8TS for $V_{DD} = 0.6V$ to $1.2V$	82
Fig. 3.15	Comparison of (a) Read power (μW) (b) Write power (μW) of 8TS and 8TD SRAM Cells as a function of supply voltage	83
Fig. 3.16	Monte Carlo (2000 point) analysis of differential 8TD and single-ended 8TS SRAM cells at $V_{DD} = 0.6V$. Butterfly curves for (a) 8TD (b) 8TS. The histograms for the SNM extracted from the butterfly curves for (c) 8TD and (d) 8TS.	84
Fig. 4.1	Circuit schematics for implementation of single bit boolean logic for two operands A,B along with the sensing strategy employed. (a) 6T SRAM-IMC (b) 8TC SRAM-IMC (c) 8TB SRAM-IMC and (d) 8TE SRAM-IMC	90
Fig. 4.2	SRAM bit lines (BL/BLB) voltage levels to be sensed by SAs for IMC bitwise logical operations for (a) 6T SRAM cell (b) 8TC SRAM cell (c) 8TB SRAM cell and (d) 8TE SRAM cell	91

Fig. 4.3	Transient response of SRAM bitcells in compute-mode for all possible bit combinations of two single bit operands (A,B) '00', '01', '10 and '11' at V_{DD} = 0.8 V for (a) 6T SRAM cell (b) 8TC SRAM cell (c) 8TB SRAM cell (d) 8TE SRAM cell	92
Fig. 4.4	Circuit schematic for implementation of single bit boolean logic (NAND/NOR and XOR/XNOR) for two operands A,B using proposed 8TD SRAM cell	94
Fig. 4.5	Transient response for compute mode of proposed 8TD SRAM cell for all possible input operand combinations (AB), '00,01,10,11' using (a) concurrent word-line and (b) sequential word-line activations at V_{DD} = 0.8 V.	95
Fig. 4.6	Transient response of the asymmetric sense amplifier for 6k Monte carlo simulations at TT corner for T = 300.15 K and V_{DD} = 0.8 V to implement (a) NAND and (b) NOR outputs for all possible input cases of input operands (AB), '00,01,10,11', at for ΔV_{TH} = 50 mV.	96
Fig. 4.7	Distribution of latency for NOR and NAND logic operation for all possible combinations of input operands (AB), '00,01,10,11' for 6k Monte Carlo simulation runs at T = 300.15 K and V_{DD} = 0.8 V with ΔV_{TH} = 50 mV.	96
Fig. 4.8	Transient simulations for (a) XNOR and (b) XOR logic operations for all possible input operand combinations, AB = 00, 01, 10, and 11. The results are obtained from 2000 Monte Carlo simulation runs at T = 300.15 K, V_{DD} = 0.8 V, and ΔV_{TH} = 50 mV.	97
Fig. 4.9	Latency distributions for (a) XNOR and (b) XOR logic operations for all possible input operand combinations, AB = 00, 01, 10, and 11. The results are obtained from 2000 Monte Carlo simulation runs at T = 300.15 K, V_{DD} = 0.8 V, and ΔV_{TH} = 50 mV.	98
Fig. 4.10	SRAM-IMC latency distributions under device mismatch at V_{DD} = 0.8 V and ΔV_{TH} = 50 mV: (a) violin-plot summary with extracted μ and σ, (b) Monte Carlo latency distribution for the fast cluster consisting of 6T, 8TB, and 8TE cells, and (c) Monte Carlo latency distribution for the higher-latency 8TC and 8TD cells.	99

Fig. 4.11	Performance comparison of SRAM-IMC cell architectures as a function of V_{DD} at the TT process corner and $T = 300.15$ K: (a) latency, (b) average energy per bit, and (c) throughput for 6T, 8TC, 8TB, 8TE, and 8TD SRAM-IMC cells.	101
Fig. 4.12	Energy reduction of the proposed 8TD SRAM-IMC cell with respect to 6T, 8TB, and 8TE SRAM-IMC macros across V_{DD}.	103
Fig. 4.13	Layout views of the SRAM cells considered for IMC implementation: (a) 8TE, (b) 6T, (c) 8TC, (d) 8TB and (e) 8TD. The layouts are used to support the normalized area comparison reported in Table 4.2.	105
Fig. 5.1	Proposed reconfigurable SRAM-IMC architecture showing the shared memory-mode and compute-mode data paths.	111
Fig. 5.2	SNM scaling of the proposed SRAM cell for 2×2 to 32×32 arrays over $V_{DD} = 0.6$–1.2 V: (a) absolute SNM and (b) relative SNM deviation, ΔSNM with respect to the 2×2 reference array.	113
Fig. 5.3	Array-size dependence of write-margin deviation, $\lvert WM - V_{DD}/2\rvert$, across the supply-voltage range.	114
Fig. 5.4	Read-delay variation with array size from 2×2 to 32×32 across the supply-voltage sweep.	116
Fig. 5.5	Read-power variation with array size from 2×2 to 32×32 across the supply-voltage sweep. The logarithmic y-axis highlights the wide power spread under array scaling.	117
Fig. 5.6	Compute-mode NAND-evaluation latency of the proposed SRAM-IMC array for 2×2 to 32×32 organizations across $V_{DD} = 0.6$–1.2 V.	120
Fig. 5.7	Average compute energy per bit of the proposed SRAM-IMC array for 2×2 to 32×32 organizations across $V_{DD} = 0.6$–1.2 V.	121
Fig. 5.8	Compute-mode throughput of the proposed SRAM-IMC array for 2×2 to 32×32 organizations across $V_{DD} = 0.6$–1.2 V.	122

CHAPTER – 1

INTRODUCTION AND LITERATURE REVIEW

1.1 OVERVIEW OF THE CHAPTER

Modern computing systems are increasingly limited by the cost of moving data between memory and logic units. This limitation becomes more critical in data-intensive workloads, where large volumes of operands must be repeatedly fetched, processed, and written back. Although SRAM is already widely used for high-speed on-chip storage because of its low access latency and CMOS compatibility, conventional SRAM still functions mainly as a passive storage block. As supply voltage scales and array dimensions increase, SRAM operation is further affected by reduced sensing margin, bit-line loading, process-induced variation, leakage, and stability degradation. These issues motivate the need for SRAM design strategies that improve throughput without sacrificing reliability and energy efficiency. In this context, SRAM-based in-memory computing has emerged as an important direction because it allows selected logic operations to be performed inside or near the memory array, thereby reducing unnecessary data movement.

To address this need, the chapter builds a circuit-to-array viewpoint for SRAM-based in-memory computing. The discussion first introduces the basic operation of SRAM arrays, conventional SRAM bit-cells, bit-line sensing, and SRAM-IMC macro-level logic operation. It then develops the literature background by reviewing representative sense amplifier architectures, SRAM bit-cell designs, SRAM-IMC macros, and SRAM-IMC array organizations. The purpose is to identify the limitations that remain when SRAM is expected to support both memory and compute operations. These limitations include sense-amplifier offset and timing sensitivity, bit-cell stability and half-select concerns, macro-level peripheral overhead, and array-level scaling effects such as increased bit-line capacitance, sensing-margin degradation, and energy growth. The solution direction adopted in this thesis is therefore based on coordinated optimization of sensing, bit-cell robustness, SRAM-IMC macro operation, and reconfigurable array-level integration.

The chapter is organized as follows. Section 1.2 introduces the fundamentals of SRAM and SRAM-based in-memory computing, including SRAM array organization, bit-cell operation, sensing, macro-level logic, and array-level reconfigurability. Section 1.3 discusses the background and motivation of the work in relation to the memory-wall problem. Section 1.4 reviews prior work on SRAM sense amplifiers, SRAM bit-cells, SRAM-IMC macros, and SRAM-IMC array architectures. Based on this review, Section 1.5 identifies the research gaps, Section 1.6 presents the problem statement, Sections 1.7 and 1.8 present the objectives and methodology, and Section 1.9 gives the thesis organization.

1.2 FUNDAMENTALS OF SRAM AND SRAM-BASED IN-MEMORY COMPUTING

Static random-access memory, commonly referred to as SRAM, is one of the most important memory structures used in digital integrated circuits. It is called *static* because the stored data remain available as long as the supply voltage is maintained. Unlike dynamic random-access memory, SRAM does not require periodic refresh. This makes it suitable for memory blocks that must be accessed frequently and quickly, such as cache memories, register files, local buffers, and embedded memory arrays. In modern processors and system-on-chip architectures, SRAM is usually not the largest memory in the system, but it is often the fastest memory available to the computing core. For this reason, improvements in SRAM speed, energy, and reliability can strongly affect the overall behavior of the system [1].

The fundamentals of SRAM and SRAM-based in-memory computing may be understood through four closely related design levels: the SRAM bit-cell, the read and sensing path, the SRAM-IMC macro, and the reconfigurable SRAM-IMC array. These levels also define the technical progression of this thesis. The following subsections introduce each level in an elementary manner before the detailed state-of-the-art review is presented.

1.2.1 SRAM ARRAY ORGANIZATION AND CONVENTIONAL MEMORY ACCESS

An SRAM array is made by arranging many SRAM bit-cells in rows and columns. The cells store the data, while the surrounding peripheral circuits allow the required data to be accessed. The main peripheral blocks are the row decoder, column selection circuitry, precharge circuit, write driver, sense amplifier, and timing-control logic. The row decoder selects the required row through a word line, while the bit-lines associated with each column carry the data during read and write operations. The complete SRAM therefore cannot be understood only by looking at the bit-cell. The cell, bit-lines, word lines, precharge circuit, and sense amplifier all work together during memory access.

Figure 1.1 shows the basic organization of an SRAM array. The memory cells are placed at the intersections of word lines and bit-line pairs. A selected word line activates all cells in one row, while the corresponding bit-line pairs carry the information stored in each selected cell. During write operation, the column drivers force the bit-lines to the required data values. During read operation, the bit-lines are first precharged and then allowed to develop a small differential voltage according to the stored data. This small differential is finally resolved by the sense amplifier. Therefore, even a simple SRAM access involves a coordinated sequence of row selection, bit-line preparation, cell access, and sensing.

1.2.2 SRAM BIT-CELL OPERATION AND STABILITY METRICS

The smallest storage element in SRAM is the bit-cell. The most common SRAM cell is the conventional 6T cell, where “6T” denotes six transistors. Four of these transistors form two cross-coupled inverters, and the remaining two transistors act as access devices. The cross-coupled inverters store the data. One internal node stores the true value Q, while the other stores the complementary value $\overline{Q}$. If Q is high, $\overline{Q}$ is low; if Q is low, $\overline{Q}$ is high. This

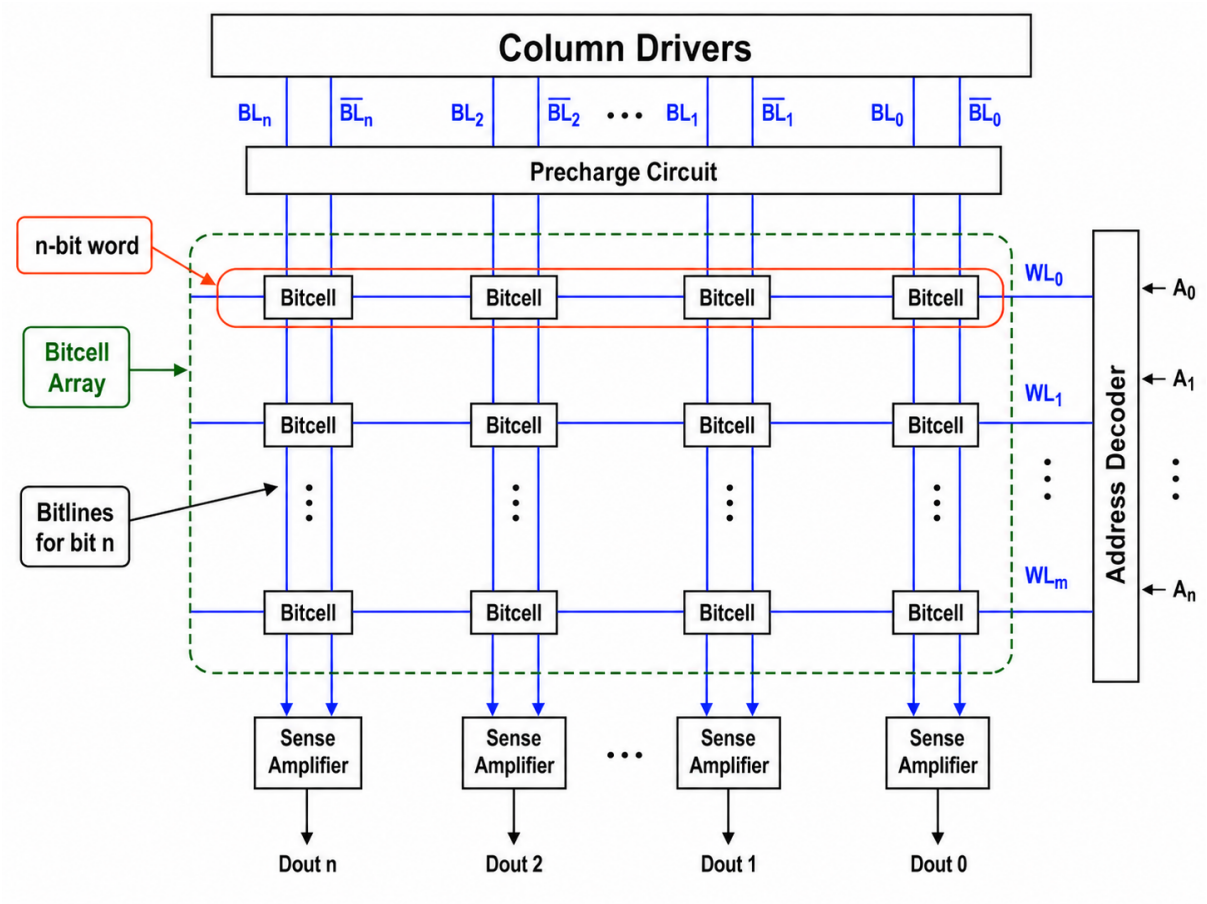


Figure 1.1: Basic organization of an SRAM array showing the bitcell matrix, wordline selection through the address decoder, differential bitline pairs, column drivers, precharge circuit, and sense-amplifier readout path.

complementary state is maintained by the feedback action of the two inverters. Thus, a bit in SRAM is stored as the stable state of a latch rather than as charge on a capacitor.

The conventional 6T SRAM bit-cell is shown in Fig. 1.2. The two access transistors connect the internal storage nodes to the complementary bit-lines BL and BLB when the word line is activated. In the hold condition, the word line remains inactive and the cross-coupled inverters retain the stored state. In this mode, the storage nodes are isolated from the bit-lines and the cell behaves as a bistable latch. This hold condition is the basic reason SRAM can retain data as long as the supply voltage is maintained.

The write operation is used to change the stored state of the cell. To write data, the write driver first forces the two bit-lines to complementary values. For example, to write a logic 1 at node Q, BL is driven high and BLB is driven low. To write a logic 0, BL is driven low and BLB is driven high. After the bit-lines are prepared, the word line is activated. This turns on the access transistors and connects the storage nodes to the bit-lines. If the bit-line drivers are strong enough to overcome the previous state of the cross-coupled inverters, the cell flips to the new value. The ease with which the cell can be overwritten is described by its write ability or write margin.

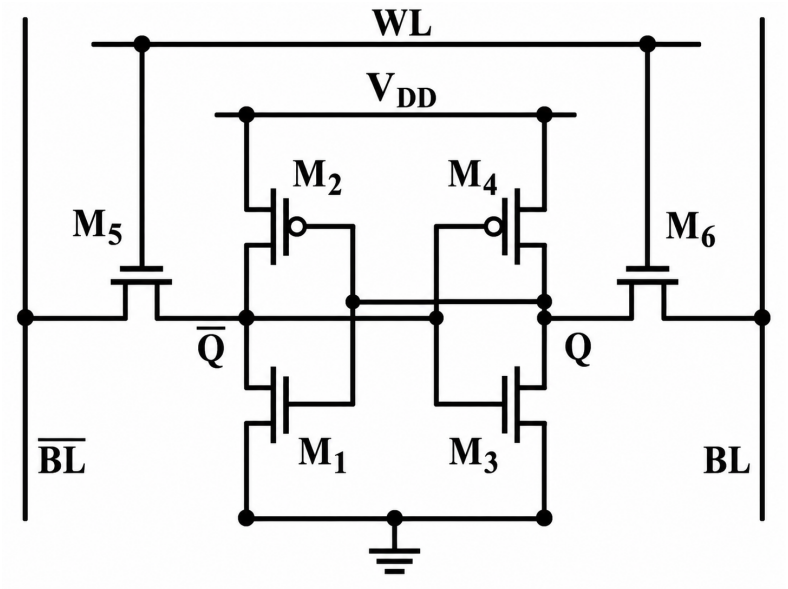


Figure 1.2: Conventional 6T SRAM bit-cell showing the storage nodes and differential bit-line access.

The read operation is more delicate because the stored value must be sensed without unintentionally changing it. Before a read operation, both bit-lines are usually precharged to V_{DD}. When the word line is activated, the cell connects its internal storage nodes to the bit-lines. One of the bit-lines begins to discharge through the pull-down path associated with the node storing logic 0, while the other bit-line remains close to the precharged level. This creates a small voltage difference between BL and BLB. The read operation is therefore based on detecting a small differential signal, not on fully discharging one of the bit-lines.

Several basic quantities are used to describe SRAM behavior. Static noise margin (SNM) measures how strongly the cell can retain its stored state in the presence of noise. Write margin describes how easily the stored value can be changed during a write operation. Read delay is the time required to access the stored data and produce a valid output. Read and write power measure the energy cost of active memory operation. Leakage power measures the power consumed even when the cell is not actively switching. These quantities are affected by supply voltage, transistor strength, mismatch, temperature, and array loading. In scaled technologies, reduced supply voltage and process-induced variation make these metrics harder to satisfy simultaneously [1, 2].

A useful first-order expression for dynamic switching energy is

$$E_{\mathrm{dyn}} \approx \alpha C_{\mathrm{sw}} V_{\mathrm{DD}}^2, \tag{1.1}$$

where C_{sw} is the effective switched capacitance and α is the switching activity factor. This expression explains why reducing V_{DD} is attractive for low-power operation. However, SRAM cannot be scaled in voltage without limit. A lower supply voltage reduces dynamic energy, but it also weakens the access devices, slows bit-line development, reduces sensing margin, degrades write ability, and increases sensitivity to threshold-voltage mismatch. Practical SRAM design is therefore a balance between energy efficiency and reliable operation.

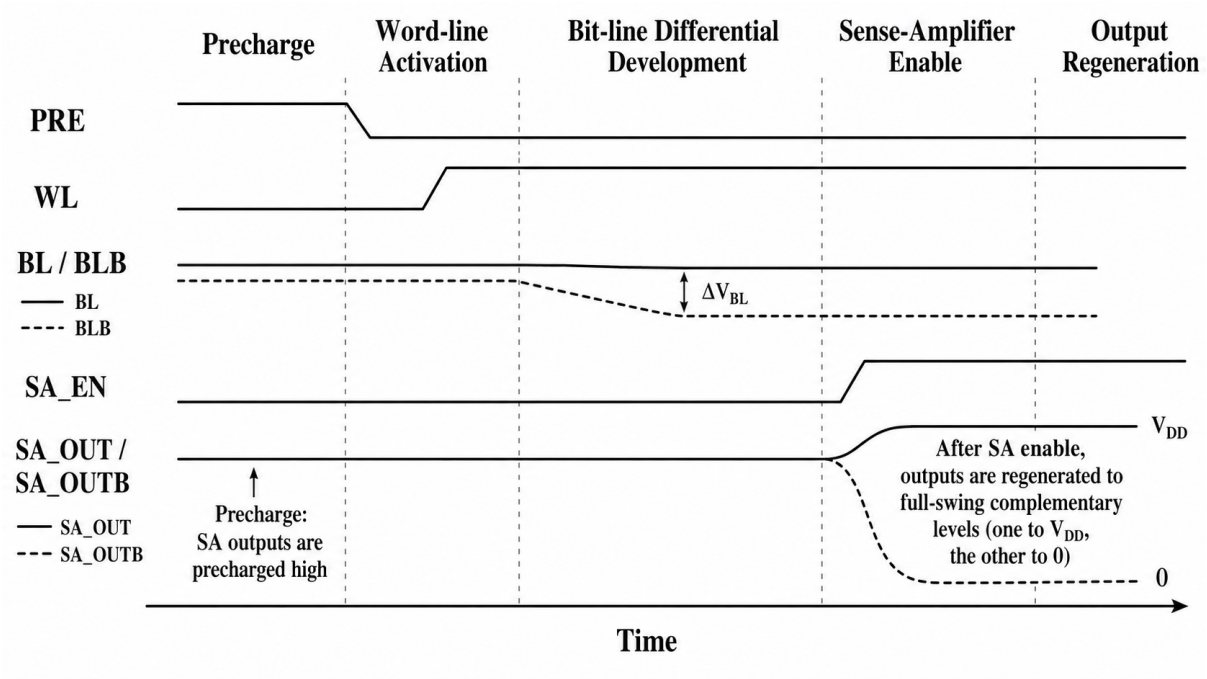


Figure 1.3: Typical SRAM read sequence showing precharge, word-line activation, bit-line differential development, sense-amplifier enable, and full-swing output regeneration.

1.2.3 SENSE-AMPLIFIER OPERATION IN SRAM READOUT

A full bit-line discharge would be slow and energy expensive because bit-lines are long, heavily loaded wires shared by many cells in a column. Therefore, practical SRAM read operation allows only a small voltage difference to develop on BL and BLB before sensing begins. This small voltage difference is then amplified by the sense amplifier. The sense amplifier is a high-gain regenerative circuit that converts the small bit-line differential into a full-swing digital output. In this way, the sense amplifier reduces read delay and energy by avoiding the need for a large bit-line swing.

Figure 1.3 illustrates the typical timing sequence of an SRAM read operation. During the precharge phase, the bit-lines are initialized to a known high voltage level. After precharge is released, the word line is activated and the selected bit-cell begins to create a small voltage difference between BL and BLB. The sense amplifier is then enabled only after sufficient bit-line differential has developed. Once enabled, the regenerative action of the sense amplifier pulls its outputs to full-swing complementary voltage levels. This sequence ensures that the cell is not required to fully discharge the bit-line, while the final output is still restored to a valid digital level.

The sense amplifier is one of the most important peripheral circuits in SRAM. Its task is simple to state but difficult to perform reliably: it must make a fast and correct decision from a very small input difference. A commonly used sense amplifier structure is based on a cross-coupled latch. When the sense amplifier is enabled, positive feedback amplifies the small imbalance at its inputs until one output is pulled high and the other is pulled low. The speed of this decision depends on the available bit-line differential, the capacitance at the sensing

nodes, transistor mismatch, and the strength of the regenerative latch. If the sense amplifier is enabled too early, the bit-line difference may be too small and the read operation may become slow or incorrect. If it is enabled too late, read delay increases because unnecessary time is spent developing a larger bit-line swing. Thus, sense-amplifier timing is a critical part of SRAM design.

Bit-line capacitance strongly affects both SRAM delay and energy. In a large array, many cells are attached to the same column bit-lines. As the number of rows increases, the effective bit-line capacitance also increases. A larger capacitance takes more time to charge or discharge and consumes more dynamic energy. Therefore, the read performance of SRAM is not only determined by the strength of the selected bit-cell; it is also governed by the total bit-line loading and the ability of the sense amplifier to resolve small voltage differences quickly. This is why SRAM read path design is usually treated as a combined problem involving the cell, the bit-line, and the sense amplifier.

1.2.4 SRAM-IMC MACRO-LEVEL LOGIC OPERATION

In a conventional computing system, SRAM stores data and separate logic circuits perform computation. During an operation, operands are read from memory, transferred to the processing unit, processed, and then written back. This organization is general and flexible, but it can be inefficient for workloads that repeatedly access large amounts of data. In such cases, the energy and delay associated with data movement may become comparable to, or even larger than, the energy and delay of the computation itself. This observation motivates near-memory computing and in-memory computing, where the goal is to reduce the distance and frequency of data movement [3, 4].

In-memory computing goes beyond simply placing logic near memory. It attempts to use the memory structure itself as part of the computing mechanism. In SRAM-based in-memory computing, selected operations are performed within, or immediately around, the SRAM array by exploiting word-line activation, bit-line voltage development, charge sharing, or sensing behavior. The memory array is still used to store data, but under specific control conditions it also participates in computation. This makes SRAM-IMC attractive because it can reuse the high internal bandwidth and column-level parallelism of the memory array.

A simple way to understand SRAM-IMC is to consider what happens when more than one stored value influences a shared bit-line. In a normal read operation, one selected cell produces a bit-line response corresponding to one stored bit. In a compute operation, two or more selected cells can interact through the same bit-line structure. The resulting bit-line voltage may then represent a function of the stored data rather than the value of a single cell. With suitable control and sensing, this behavior can be used to implement Boolean operations such as NAND, NOR, XOR, and XNOR [5, 6].

At the macro level, SRAM-IMC combines a memory cell array with compute-aware peripheral circuits. A macro is larger than an isolated bit-cell but smaller than a complete processor or memory subsystem. It is the block where the memory cells, word-line drivers, bit-lines, precharge

devices, sense amplifiers, and compute logic are brought together to perform a specific function. In Boolean SRAM-IMC, the operands are stored inside SRAM cells. The compute operation is performed by activating the relevant control signals and interpreting the resulting bit-line response through the sensing path.

The distinction between memory mode and compute mode is important. In memory mode, the SRAM array behaves like a conventional memory. One row is selected, the stored bit develops a signal on the bit-lines, and the sense amplifier produces the data output. In compute mode, the same physical structure is used differently. The word-line activation pattern, bit-line interaction, or sensing interpretation is modified so that the output corresponds to a logic function of stored operands. Thus, the same array can act as a memory in one mode and as a simple computing fabric in another mode.

SRAM-IMC can be implemented in digital, analog, or mixed-signal forms. Digital SRAM-IMC preserves discrete logic interpretation and is commonly used for Boolean logic, bitwise operations, search, and low-precision digital computation. Analog or mixed-signal SRAM-IMC uses current summation, voltage accumulation, capacitive charge sharing, or other analog effects to perform operations such as dot products and multiply–accumulate functions. Digital SRAM-IMC is generally easier to verify and more robust against small analog errors, whereas mixed-signal SRAM-IMC can exploit stronger parallelism but is more sensitive to noise, mismatch, reference generation, and conversion overhead [4, 7].

Throughput in SRAM-IMC depends on both speed and parallelism. A simple expression is

$$\text{Throughput} \approx \frac{N_{\text{op}}}{t_{\text{op}}}, \tag{1.2}$$

where N_{op} is the number of operations completed during one evaluation and t_{op} is the evaluation time. In a conventional logic unit, parallelism is limited by the number of available processing lanes. In SRAM-IMC, many columns of the array can potentially operate in parallel, allowing multiple bit-level operations to be evaluated together. Therefore, SRAM-IMC can improve effective throughput by reducing data movement and by exploiting the natural parallel structure of the memory array.

1.2.5 ARRAY-LEVEL SRAM-IMC OPERATION AND RECONFIGURABILITY

SRAM-IMC introduces design challenges that become more severe as the computation is moved from a small macro to a larger array. If multiple cells are activated together, the internal storage nodes may be disturbed, especially in cells where the read and write paths are not separated. The bit-line voltage levels may also become closer together, making the sensing decision more difficult. Additional compute circuitry may increase area and energy overhead. In analog or mixed-signal IMC, the output may depend strongly on device matching, reference voltages, and peripheral conversion accuracy. Thus, a useful SRAM-IMC design must not only compute correctly; it must also preserve the normal memory function and remain robust under variation.

The SRAM bit-cell plays a central role in this trade-off. A conventional 6T cell is compact and fast, but because it uses the same access path for read and write, it is more vulnerable to read

disturbance during multi-row or compute-style activation. Read-decoupled cells, differential cells, and multi-transistor cells can reduce this risk by separating the storage nodes from the read or compute path. Such cells may improve robustness and sensing margin, but they also add area and control complexity. Therefore, the bit-cell selected for SRAM-IMC must be judged not only by its standalone SNM or write margin, but also by how safely it supports compute-mode access.

The sense amplifier is equally important in SRAM-IMC. In a normal read operation, the sense amplifier decides whether the selected cell stores 0 or 1. In compute mode, the sense amplifier may need to distinguish among bit-line responses produced by different operand combinations. For example, the responses for AB = 00, 01, 10, and 11 may have to be separated reliably. If the sense amplifier has large offset, poor timing control, or insufficient sensitivity, the compute output may become incorrect even if the cell operation is otherwise valid. Thus, in SRAM-IMC, sensing correctness becomes computation correctness.

At the array level, the problem becomes broader. A small SRAM-IMC macro may demonstrate correct logic operation, but a larger array introduces longer bit-lines, higher capacitance, larger interconnect loading, more complex precharge requirements, and stronger variation sensitivity [7, 8]. These effects influence normal memory operation as well as compute operation. In memory mode, they affect SNM, write margin, read delay, and read power. In compute mode, they affect logic-evaluation latency, average energy per bit, throughput, and output separability. Hence, array-level evaluation is necessary before a proposed SRAM-IMC technique can be considered scalable.

A reconfigurable SRAM-IMC array is designed to operate in both memory and compute modes. In memory mode, it behaves like a conventional SRAM array and supports normal data storage and readout. In compute mode, it reuses the stored data, bit-lines, and sensing path to perform selected logic operations. The benefit of such an array is that the same physical memory can serve two roles. The challenge is that the compute mode must be added without seriously degrading the memory mode. If the additional compute circuitry causes large area, power, or reliability penalties, the advantage of in-memory computing is reduced.

1.2.6 SRAM-TO-IMC DESIGN HIERARCHY

For this reason, SRAM-IMC design is best understood as a hierarchy. At the lowest level, the bit-cell must store data reliably. At the read-path level, the bit-lines and sense amplifier must convert the stored information into a valid output. At the macro level, selected cells and peripheral circuits must be arranged so that stored operands can participate in logic evaluation. At the array level, the same idea must remain stable, energy efficient, and scalable across many rows and columns. Each level affects the next. A weak bit-cell limits the macro. A poor sense amplifier limits both read and compute correctness. A compute macro that works only for a small number of cells may not scale to a useful array.

This thesis follows this hierarchy. The sense amplifier is first studied because fast and reliable sensing is essential for both SRAM read and SRAM-IMC output generation. The SRAM bit-cell is then studied because storage stability, write ability, and disturbance immunity originate at the

cell level. The SRAM-IMC macro is then examined to show how the bit-cell and sensing strategy can be combined to perform Boolean logic inside the memory structure. Finally, the work is extended to an SRAM-IMC array to evaluate whether the proposed memory-compute approach remains useful when operated in both memory and compute modes under array scaling.

The purpose of this section has been to establish the basic working principles required for the rest of the chapter. Conventional SRAM operation depends on the coordinated interaction among the bit-cell, bit-lines, precharge circuit, decoder, and sense amplifier. SRAM-IMC modifies this interaction so that the memory array can participate in computation as well as storage. With this foundation, the next sections review prior work on SRAM sense amplifiers, SRAM bit-cells, SRAM-IMC macros, and SRAM-IMC arrays, and then identify the research gaps addressed in this thesis.

1.3 BACKGROUND AND MOTIVATION

Modern signal- and data-processing workloads impose stringent latency and energy constraints, particularly in AI-driven applications deployed across cloud and edge systems. CNN and DNN inference pipelines repeatedly move activations and weights between memory and compute units, and this movement often dominates overall energy and latency more than arithmetic operations themselves. Recent CNN-focused accelerator studies have reinforced this trend [9]. For deep neural network deployment in broader settings, software-hardware co-design analyses similarly identify data movement as a primary bottleneck [10, 11]. As model complexity, precision requirements, and deployment scale continue to increase, this imbalance becomes a first-order system bottleneck, commonly referred to as the memory wall or von Neumann bottleneck.

This architectural progression can be understood in three stages. In the conventional von

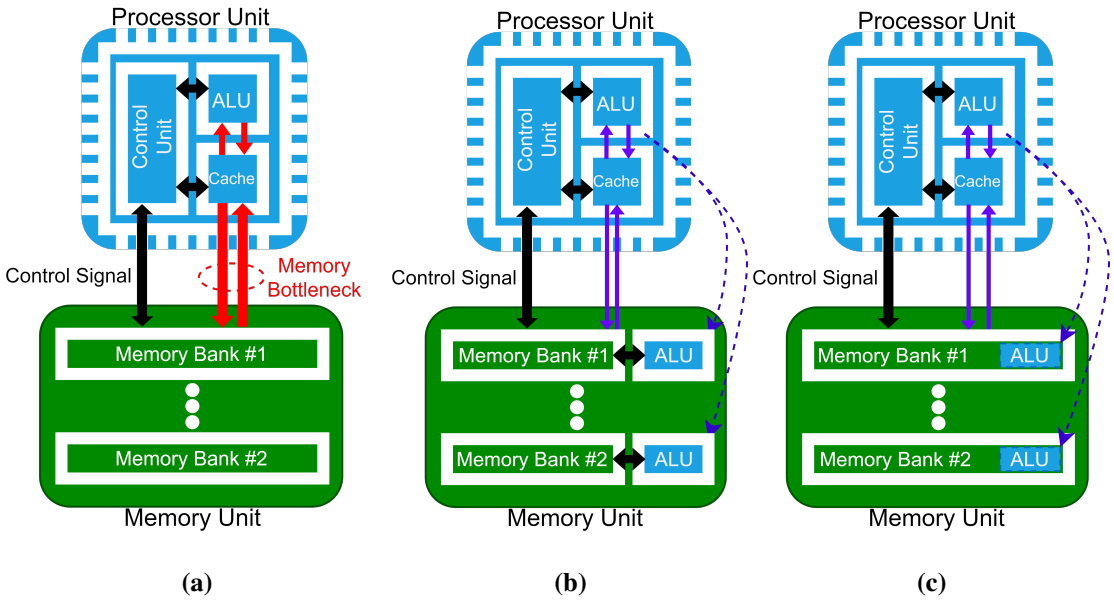


Figure 1.4: Memory bottleneck dilemma and architectural alternatives: (a) conventional von Neumann organization, (b) near-memory computing, (c) in-memory computing.

Neumann organization, computation and memory remain physically separated, forcing continuous round-trip data movement through shared interfaces and creating a bandwidth-energy bottleneck. Near-memory computing places compute logic closer to memory, reducing transfer distance and interface overhead while still preserving a partial boundary between storage and computation. In-memory computing goes a step further by moving selected operations directly into the memory fabric, thereby minimizing data movement further and enabling higher parallelism for memory-bound kernels. Together, these architectural directions show why memory-compute co-design is now central to improving end-to-end throughput and efficiency in modern AI-oriented systems.

SRAM remains central to high-speed on-chip storage because it provides very low access latency and, unlike DRAM, does not require periodic refresh, making it inherently better suited to latency-critical memory functions within active computing systems [12]. In modern processors, cache memory plays a decisive role in reducing effective data access time by retaining frequently used data close to the CPU, and conventional caches continue to rely on SRAM because of its fast response characteristics [13]. Beyond speed alone, SRAM has remained highly relevant due to its technological maturity, robust operation, and compatibility with advanced CMOS logic processes, all of which make it attractive for seamless integration into mainstream computing platforms and embedded memory hierarchies [4]. This continued reliance is also evident from ongoing state-of-the-art process development, where dedicated SRAM demonstrations at leading nodes such as 2 nm still receive major design focus for achieving high density and energy-efficient compute support [14]. Therefore, SRAM is not merely a legacy storage element in the memory hierarchy; rather, it continues to serve as the practical and architectural foundation for fast on-chip storage, while simultaneously providing a natural substrate for emerging SRAM-based in-memory computing paradigms that aim to reduce data movement without departing radically from standard processor-memory organizations [4, 12].

SRAM-based in-memory computing (SRAM-IMC) has gained significant attention as a practical means of alleviating the data-transfer burden inherent to conventional processor–memory organizations. In standard computing systems, even relatively simple operations may require operands to be repeatedly fetched from memory, processed in separate logic units, and then written back, causing substantial overhead in time and energy. SRAM-IMC seeks to relax this dependence on external data movement by allowing selected computations to be executed within, or immediately around, the SRAM array itself. Such computation is commonly enabled through controlled word-line activation, bit-line voltage evolution, charge sharing, and lightweight peripheral circuits that reinterpret memory access events as computational primitives [3, 4]. As a result, SRAM-IMC offers the possibility of reducing memory traffic, shortening effective execution latency, and improving energy efficiency, particularly for workloads dominated by repeated vector operations, bitwise manipulation, and multiply–accumulate patterns. Its appeal is further strengthened by the fact that SRAM is already deeply embedded in modern on-chip memory hierarchies, making SRAM-IMC more compatible with existing digital systems than many alternative in-memory platforms. Recent 2026 works also suggest that SRAM-IMC is progressing beyond conceptual demonstrations toward more specialized and configurable

realizations, including fully digital compute-in-memory macros for hybrid SNN–CNN edge workloads, ternary and Boolean SRAM-CIM macros, multimode segmented-charge-sharing MAC-oriented designs, and compact high-precision 10T SRAM-CIM implementations [15–18].

Although SRAM-IMC has shown strong potential, the problem is still far from fully solved in practice. Existing SRAM-IMC realizations continue to face important limitations in sensing reliability, read/write stability, and robustness of analog or mixed-signal computation, especially when precision requirements increase and operating conditions become more constrained [7, 19]. These difficulties are closely tied to narrow bitline-voltage margins, charge-sharing instability, process-sensitive behavior, and the non-ideal effects introduced by multirow activation and peripheral support circuits, all of which can degrade computational accuracy and restrict reliable large-scale deployment [7, 19]. In addition, practical macro-level designs must still balance stability, energy, area efficiency, and variation tolerance at the array level, where read/write margin degradation and Monte Carlo variability remain significant concerns even in improved 10T SRAM-IMC structures [20].

Taken together, the above discussion shows that meaningful gains in SRAM-IMC cannot be achieved through isolated optimization of a single circuit block. Instead, robust throughput improvement requires coordinated attention to sensing behavior, bitcell stability, macro organization, and array-level scalability under realistic operating and variation constraints. This systems-to-circuit viewpoint defines the central thesis direction and motivates the detailed technical review that follows.

1.4 LITERATURE REVIEW

The literature on SRAM and SRAM-based in-memory computing spans several closely connected design layers. Improvements at one layer often influence the behavior of the remaining layers; for example, a faster sense amplifier can improve read delay but may still be limited by insufficient bitline differential, while a robust bitcell may lose its advantage if array-level loading or peripheral overhead becomes dominant. For this reason, the review cannot be restricted to only SRAM-IMC macro architectures. It must also consider the supporting SRAM blocks that determine whether in-memory computation can be performed reliably and efficiently. The present review therefore begins with SRAM sense amplifiers, since sensing speed, offset tolerance, and low-voltage decision capability directly affect both memory and compute outputs. It then reviews SRAM bitcell architectures, with emphasis on stability, writability, half-select immunity, leakage, and variation tolerance. The discussion is further extended to SRAM-IMC macros, where the interaction between stored data, wordline control, bitline behavior, and sensing circuitry is used to perform logic or arithmetic functions. Finally, SRAM-IMC array architectures are reviewed to examine how macro-level ideas scale when array capacitance, peripheral sharing, precision requirements, and reconfigurability are considered. This organization provides the basis for identifying the research gaps addressed in the thesis.

1.4.1 SENSE AMPLIFIER ARCHITECTURES

The sense amplifier (SA) is one of the most critical peripheral circuits in SRAM because it determines how rapidly and reliably the small differential signal developed on the complementary bit-lines can be converted into a full-swing digital read output. During a read operation, the storage cell produces only a limited voltage or current difference between the bit-lines, while the large bit-line capacitance, reduced supply voltage, and process-induced device mismatch make this signal increasingly difficult to detect with high speed and accuracy. As a result, the SA has a direct impact on read delay, energy consumption, read yield, and low-voltage robustness of the overall memory system. With technology scaling, these challenges become more pronounced due to smaller sensing margins and greater threshold-voltage variation, making the design of an efficient and variation-tolerant SA a key requirement in modern SRAM architectures. Accordingly, a wide range of SRAM SA topologies have been developed to improve sensing speed, reduce power, enhance stability, and maintain reliable operation under low-voltage and parameter-variation conditions.

Beyond the conventional voltage-mode and current-mode families, SRAM sense amplifier design has progressively moved toward assisted, variation-aware, and application-specific architectures intended to sustain reliable operation under scaled supply voltages and technology-induced mismatch. In these designs, the primary emphasis shifts from only improving raw sensing speed to also mitigating offset, enhancing read yield, improving low-voltage robustness, and in some cases enabling more flexible read/compute functionality. Representative examples include redundancy-based sensing, reference-assisted operation, capacitive or delay-aware offset correction, body-bias based offset mitigation, reconfigurable sensing paths, and more recent supply-boosted or ultra-low-swing latch-based schemes [21–28]. To place these developments in a broader historical and architectural context, Table 1.1 summarizes representative SRAM sense amplifier architectures reported over time along with their key design techniques and associated trade-offs.

Table 1.1: Representative sense amplifier architectures and key trade-offs in design.

Year	Ref.	Technique	Observations
1976	Heller et al. [29]	Charge-Transfer SA	• Foundational charge-transfer sensing • Fast response with reduced input offset • Less dependence on large bitline swing • Requires intermediate precharge/bias conditions
1991	Blalock & Jaeger [30]	Clamped Bitline SA	• Current-mode sensing with reduced bitline swing • Dedicated clamp transistors on bitlines • Improved stability for small signals • Area and power overhead from clamps
1992	Chee et al. [31]	Hybrid Current SA	• Current-mode architecture with latch • Additional current conveyor stage • Reduced bitline noise coupling • Increased device count and biasing

Year	Ref.	Technique	Observations
1993	Kobayashi et al. [32]	Current-Controlled Latch SA	• Voltage-mode latch-based sensing • High input impedance from bitlines • Simple and compact architecture • Sensitive to mismatch and offset
2001	Khellah & Elmasry [33]	Modified Clamped SA	• Current-mode sensing with precharge • Extra precharge and equalization devices • Improved bitline symmetry • Added area and power overhead
2004	Wicht et al. [34]	Decoupled Latch SA	• Voltage-mode latch with input decoupling • Reduced kickback noise • Faster sensing operation • Offset tolerance not addressed
2007	Schinkel et al. [35]	Double-Tail SA	• Two-stage voltage-mode sensing • Separate tail devices for isolation • Improved speed–power trade-off • Higher transistor count
2007	Do et al. [36]	Modified Hybrid SA	• Hybrid sensing using conveyor and latch • Additional BL–DL coupling devices • Improved sensing reliability • Increased control and biasing
2007	Verma & Chandrakasan [37]	Sense-Amplifier Redundancy	• Parallel redundant sensing paths • Improves sub-threshold read robustness • Reduces mismatch-induced failures • Area overhead from duplicated SAs
2010	Kawasumi et al. [38]	HCI-Trimmed SA	• Offset reduction using trimming • Supports low-supply SRAM operation • Calibration-based robustness improvement • High-voltage trimming and aging concerns
2011	Sharma et al. [21]	Redundancy SA	• Multiple parallel sense amplifiers • Statistical reduction of mismatch errors • Improved read yield • Area and energy duplication
2012	Sharma et al. [39]	Charge-Limited Sequential SA	• Ultra-low-energy sensing approach • Low bitline precharge voltage • Strong energy reduction potential • Larger delay and extra control complexity
2016	Wen et al. [22]	Reference-Based SA	• Differential sensing with reference levels • Reference generation circuitry required • Improved low-voltage robustness • Sensitive to reference tracking
2019	Fragasse et al. [23]	Offset-Corrected SA	• Capacitive offset compensation • Replica-assisted self-timing • Improved sensing accuracy • Calibration overhead

Year	Ref.	Technique	Observations
2019	Patel et al. [40]	Hybrid Latch-Type Offset Tolerant SA	• Hybrid latch-based offset-tolerant sensing • Reduced input-referred offset • Improved low-voltage robustness • Added design complexity for offset control
2019	Reniwal et al. [41]	Auto-Calibrated SA	• Offset prediction and calibration • Improved energy-efficient sensing • Better mismatch tolerance • Calibration circuitry overhead
2021	Patel et al. [24]	Body-Biased SA	• Offset mitigation using body bias • Enhanced low-voltage read yield • Requires bias generation circuits • Sensitive to bias stability
2021	Chen et al. [25]	Reconfigurable Compute SA	• Reconfigurable read and compute paths • Single-ended and differential modes • Enables compute-in-memory • Higher architectural complexity
2022	Dutt et al. [42]	Voltage-Mode SA	• High-performance low-power voltage sensing • Cross-coupled latch-based structure • Improved delay-power balance • Conventional offset limitation remains
2022	Zhao et al. [26]	Offset-Cancellation SA	• Delay-aware offset cancellation • Additional timing control devices • Improved variation robustness • Extra sensing latency
2023	Shen et al. [43]	Single-Ended Offset-Canceling SA	• Single-ended wide-voltage sensing • Offset-canceling operation • Supports broad supply range • Extra cancellation/control overhead
2024	Sharma et al. [27]	Supply-Boosted Voltage SA	• Supply-boosted sensing scheme • MOSCAP-based charge reuse • Enhanced read speed • Minimal structural modification
2025	Kim et al. [28]	Asymmetric Voltage Latch & Ultra-Low-Swing SA	• Asymmetric latch architecture • Ultra-low bitline swing operation • Offset reduction through capacitive coupling • Energy-efficient FinFET-compatible design

As summarized in Table 1.1, SRAM sense amplifier research has developed through a broad sequence of architectural directions rather than a single linear design path. Early work established the fundamental sensing paradigms through charge-transfer, voltage-mode, and current-mode approaches, each offering a different balance between sensing mechanism, circuit simplicity, and bit-line dependence [29, 30, 32]. Subsequent designs refined these basic paradigms through current control, decoupling, tail-based isolation, hybrid current transport, and modified clamped structures in order to improve regeneration behavior and adapt sensing operation to larger bit-line

loading and reduced supply voltages [31, 33–36]. In later years, the design space expanded further to include redundancy, calibration, reference-assisted operation, body-bias control, offset-correction, and reconfigurable or function-aware sensing, indicating that the role of the SA has progressively moved beyond simple read amplification toward robust and application-conscious memory interfacing [21, 23–26]. More recent architectures such as supply-boosted and ultra-low-swing latch-based schemes show that SRAM sensing continues to be actively revisited even in mature design spaces, particularly for low-power and high-density embedded memory applications [27, 28]. Overall, the literature reflects a gradual expansion from basic sensing functionality toward increasingly specialized architectures designed to accommodate scaling, low-voltage operation, and evolving memory-system demands.

1.4.2 KEY OBSERVATIONS FROM PRIOR SENSE AMPLIFIER DESIGNS

Although the literature reports a wide variety of SRAM sense amplifier architectures, the overall design problem cannot yet be regarded as fully resolved when the practical requirements of modern SRAM read operation are considered together. A first recurring limitation is that many architectures improve sensing speed by relaxing the dependence on large bit-line voltage development, but this advantage is often obtained through additional structural elements such as clamping devices, current-conveyor stages, equalization transistors, or staged regeneration paths, which increase design complexity and may introduce extra biasing or timing requirements [30, 31, 33, 36]. By contrast, voltage-mode latch-based schemes remain attractive because of their relatively compact structure, ease of integration, and strong regenerative behavior, yet their sensing action is still closely tied to the available bit-line differential, the capacitive loading of the bit-lines, and the offset sensitivity of the latch itself [32, 34, 35]. As a result, improvements in one direction are frequently accompanied by compromises in another, and the literature repeatedly reflects a balance rather than a complete solution.

A second important observation is that low-voltage operation and process variation have gradually shifted SA design from a purely speed-oriented problem to a robustness-oriented one. As supply voltage is reduced, the differential signal available for sensing becomes smaller, while threshold-voltage mismatch and offset become more influential in determining read correctness. This has motivated a large body of work on assisted and variation-aware sensing, including redundancy-based schemes, reference-assisted methods, trimming approaches, body-bias techniques, and delay-aware offset-cancellation mechanisms [21–24, 26, 37, 38, 40, 41]. These methods clearly strengthen read reliability and improve tolerance to mismatch or reduced sensing margin; however, they also demonstrate that robustness is seldom obtained for free. In many cases, the improvement is accompanied by additional circuitry, calibration effort, larger area, increased sensing latency, or more involved peripheral support, which may reduce the practicality of the design in dense SRAM arrays where one sense amplifier serves an entire column.

A third requirement emerging from the recent literature is that sense amplifiers are now being expected to satisfy broader functional requirements than conventional readout alone.

Reconfigurable compute-aware sensing, supply-boosted operation, and ultra-low-swing schemes indicate that the SA is increasingly being treated as an adaptive interface between the memory array and higher-level system functionality rather than merely a terminal regenerative latch [25, 27, 28, 43]. While this trend is important and technically promising, it also makes the design objective more demanding, since the SA must now be judged not only by delay or power but also by robustness, controllability, and implementation overhead in realistic operating conditions. Therefore, the literature suggests that the major unresolved issue in SRAM sense amplifier research is not the absence of fast, low-power, or offset-aware techniques in isolation, but the lack of a sufficiently balanced architecture that can simultaneously support high-speed sensing, low-voltage operation, resilience to variation, and practical circuit overhead for differential SRAM read operation.

1.4.3 ADVANCES IN SRAM BITCELL ARCHITECTURES

The SRAM bitcell is the core functional unit of the memory array, and its design directly influences read stability, write ability, leakage power, access delay, area efficiency, and the minimum reliable operating voltage. As CMOS technology scales and supply voltage is reduced, these requirements become increasingly difficult to satisfy simultaneously, since shrinking noise margins, stronger process sensitivity, and tighter energy budgets expose the limitations of conventional density-oriented cells. Consequently, SRAM bitcell research has evolved beyond the standard 6T baseline toward a wide range of modified architectures, including read-decoupled, write-assist, single-ended, Schmitt-trigger-based, disturb-free, and variation-aware cells, each developed to improve selected aspects of stability, robustness, or energy efficiency under constrained operating conditions.

Table 1.2: Evolution of SRAM cell architectures from conventional 6T designs to multi-transistor cells.

Year	Reference	Technique	Observations
2005	Chang et al. [44]	Stable 8T SRAM Cell	• Read-path isolation from storage nodes • Stronger stability in scaled nodes • Foundational 8T read-decoupled direction • Area overhead over 6T
2006	Takeda et al. [45]	Read-SNM-Free 7T SRAM	• Read-disturb-free cell philosophy • Improved low-V_{DD} read robustness • High-speed low-voltage operation • Extra device and control overhead
2007	Aly & Bayoumi [46]	Low-power 7T SRAM	• Read-path decoupling transistor • Improved stability vs 6T • Cache-oriented low power • Area overhead trade-off
2007	Kulkarni et al. [47]	Schmitt-trigger sub-threshold SRAM	• Schmitt-trigger inverter pair • Reliable operation at ~160 mV • Strong variation tolerance • Increased transistor count

Year	Reference	Technique	Observations
2008	Lin et al. [48]	Highly stable nanometer SRAM	• Stability-first scaled SRAM • Low-power operation focus • Variability-aware design • Density not prioritized
2012	Kulkarni & Roy [49]	Variation-tolerant ST SRAM	• ST-based read stability • Subthreshold voltage target • Yield-oriented robustness • Assist circuitry overhead
2017	Rajput et al. [50]	Schmitt-trigger 10T SRAM	• ST-based stability boost • Improved variation tolerance • Low-voltage operation • Higher transistor count
2017	Liu et al. [51]	TFET-based 7T SRAM	• TFET leakage reduction • Ultra-low voltage target • Energy-efficient device • Technology dependence
2018	Pasandi & Pedram [52]	Disturb-free SRAM schemes	• Read-before-write logic • Half-select disturb mitigation • Array-level reliability • Control/latency cost
2018	Sanapala et al. [53]	ST-based single-ended 7T SRAM	• ST-assisted stability • Single-ended operation • IoT-oriented low power • Area overhead
2018	Giterman et al. [54]	Secure symmetrical 8T SRAM	• Symmetrical 8T structure • Leakage-attack resilience • Security-driven robustness • Area penalty
2019	Pal et al. [55]	Write-assist 9T SRAM	• Half-select free write • Integrated write assist • Improved writability • Extra assist devices
2019	Ahmad et al. [56]	TFET-based robust 7T SRAM	• TFET leakage suppression • Low-voltage robustness • Energy-efficient operation • Process dependence
2019	Jiang et al. [57]	Latch-reinforced 10T/12T SRAM	• Quadruple latch feedback • High reliability focus • Strong read stability • Large cell area
2019	Surana & Mekie [58]	Energy-efficient 6T SRAM	• Modified single-ended read • Reduced read energy • Multimedia-oriented design • Sensing complexity

Year	Reference	Technique	Observations
2019	He et al. [59]	Disturb-free 11T SRAM	• Built-in read/write assist • Half-select elimination • Ultra-low voltage target • High transistor count
2020	Asli & Taghipour [60]	Asymmetric Fin-FET SRAM	• Back-gate controlled core • Variation-aware FinFET • Improved stability margins • Control overhead
2020	Cho et al. [61]	ST-based 9T SRAM	• One-sided ST inverter • Near-threshold operation • Enhanced read stability • Added devices
2020	Eslami et al. [62]	Low-leakage 10T SRAM	• Reduced standby leakage • Single-ended structure • High-yield focus • Area overhead
2021	Shakouri et al. [63]	High-yield 10T SRAM	• Single-ended design • Yield-oriented stability • Low standby power • Increased complexity
2022	Chen et al. [64]	Reliable 8T SRAM for LiM/CIM	• 8T read-decoupled bit-cell • Compute-in-memory support • High-speed search capability • Area overhead
2022	Abbasian [65]	Near-Threshold 10T SRAM	• 10T stability-focused architecture • Near-threshold operation • Improved energy efficiency • Increased transistor count
2023	Abbasian & Sofi-mowloodi [66]	Soft-Error-Immune SRAM Cell	• Multi-bit soft-error resilience • Low-power reliability emphasis • Fault-tolerant storage focus • Area and circuit overhead
2023	Sangalang et al. [67]	Asymmetrical Schmitt-Trigger-Based SRAM	• Asymmetrical ST storage nodes • Improved hold/write robustness • Low-power standby via voltage mode • Speed compensation / boost assist
2025	Lou et al. [68]	8T Compute-SRAM Bitcell	• Compute-in-memory compatible • Area-efficient 8T architecture • Digital CIM workloads • Design complexity
2025	Dinesh et al. [69]	8T SRAM with Dynamic Body Biasing	• SNM-free read-port organization • Dynamic body-bias for subthreshold speed • Faster write operation focus • Higher write-power trade-off

Year	Reference	Technique	Observations
2025	Srivastav et al. [70]	Feedback-Cutting Low-Power 11T SRAM	• Single-bitline feedback-cutting design • LPWAN-oriented low-power operation • Strong RSNM/WSNM and low leakage • Larger transistor count
2026	Paul et al. [71]	Soft-Error-Resilient 13T SRAM	• Radiation-prone environment target • Strong upset-resilience emphasis • FinFET-based reliability design • Area and complexity overhead

As summarized in Table 1.2, the development of SRAM bitcell architectures has progressed through a broad sequence of design directions, each aimed at improving a different aspect of stability, writability, leakage, or low-voltage reliability. Early departures from the conventional density-oriented baseline introduced 7T and 8T cells that selectively modified the feedback path or separated the read path from the storage nodes, thereby reducing read disturbance and improving write controllability without completely abandoning compact array integration [44–46]. A second major direction centered on Schmitt-trigger-based cells, where reinforced inverter characteristics were used to improve noise immunity and variation tolerance, especially in subthreshold and near-threshold regimes [47, 49, 50, 53, 61, 67]. In parallel, TFET-based, asymmetric, and single-ended architectures explored alternative device- and circuit-level strategies for lowering leakage, reducing bitline activity, and sustaining operation at reduced supply voltages [51, 56, 60, 62, 63, 69]. Another important stream of work focused on disturb-free, half-select-resilient, and assist-enhanced cells, where designers introduced read-before-write logic, built-in assist mechanisms, latch reinforcement, or feedback cutting to improve robustness in increasingly variation-sensitive operating conditions [52, 55, 57, 59, 70]. More recent bitcell designs further show that SRAM cells are now being adapted not only for conventional storage but also for security-aware and compute-oriented roles, as reflected in symmetrical secure 8T cells, soft-error-resilient multi-transistor cells, and 8T bitcells tailored for logic-in-memory or compute-in-memory operation [54, 64, 66, 68, 71]. Taken together, the literature indicates that SRAM bitcell design has evolved from a primarily density-driven problem into a broader co-optimization problem involving read stability, write ability, leakage, soft-error resilience, half-select immunity, and application compatibility under scaled and low-voltage conditions.

In addition to transistor-level modifications within the SRAM bitcell, a parallel body of work has examined how array organization, peripheral support, and statistical design choices affect the overall memory behavior. At the array level, issues such as sensing speed, yield degradation, bitline loading, leakage management, near-threshold operation, and reliability under environmental stress become increasingly important because the interaction among thousands of cells, shared wordlines and bitlines, and common peripheral circuits can amplify limitations that are not always evident from single-cell analysis alone. Representative array-level architectures and optimization strategies reported in the literature are summarized in Table 1.3.

Table 1.3: Representative SRAM Array-Level Architectures and Optimization Techniques

Year	Reference	Technique	Observations
1993	Seki et al. [72]	1-Mb CMOS SRAM array with latched SA	• 256k × 4 bit organization • Latched sense amplifier • Write equalization scheme • 6 ns access time
2005	Mukhopadhyay et al. [2]	Yield-aware SRAM array modeling	• Failure probability modeling • Yield-driven design framework • Access-time estimation • V_{DD} & temperature sensitivity
2013	Garg & Kim [8]	Energy-efficient SRAM array structures	• Array-level energy analysis • Parameter-based modeling • Mathematical optimization • Energy-minimized design space
2015	Pasandi & Fakhraie [73]	Near-threshold 9T SRAM array	• 256 kb array implementation • 1k cells per bitline • Stacked inverter structure • 0.3 V operation
2015	Frustaci et al. [74]	Error-tolerant SRAM array	• Dynamic energy–quality scaling • Voltage-scalable architecture • Bit-level optimization • Error-resilient operation
2021	Wang et al. [75]	Voltage-stacked SRAM array	• Voltage stacking power delivery • Ultra-low leakage focus • 8T zigzag cell usage • High-speed control logic
2024	Wei et al. [76]	Dual-Direction SRAM In-Array CAM	• Dual-direction logic/CAM operation • Four configurable operation modes • SRAM–logic mode compatibility • Additional control overhead
2025	Surendranathan et al. [77]	Radiation-Aware SRAM Array Analysis	• Total ionizing dose impact study • Data imprinting characterization • Node-scaling vulnerability trends • Reliability and PUF implications

As summarized in Table 1.3, prior SRAM array-level research has evolved from fast dense array implementations and yield-aware statistical modeling toward broader optimization goals involving energy efficiency, near-threshold operation, architectural error tolerance, leakage reduction, and reliability under non-ideal operating conditions. Early work emphasized practical dense-array realization and access-speed improvement, whereas later studies increasingly addressed statistical yield loss, voltage scaling, and energy-aware organization through mathematical optimization and architecture-level restructuring [2, 8, 72]. More recent designs further show that array-level concerns are no longer limited to storage density and access time alone, but also include ultra-low-leakage operation, compatibility with logic-in-memory or CAM functionality,

and robustness against radiation-induced or environment-dependent reliability degradation [73–77]. This broader progression indicates that SRAM design at the array level has become a multidimensional problem in which cell choice, peripheral circuitry, operating voltage, and reliability constraints must be considered together. These observations naturally motivate a more critical examination of how effectively prior array-oriented approaches balance performance, robustness, and implementation overhead.

1.4.4 KEY OBSERVATIONS FROM PRIOR SRAM BITCELL DESIGNS

Although the literature presents a wide range of SRAM bitcell architectures, the reported designs generally strengthen selected performance attributes rather than providing a uniformly balanced solution across all requirements of scaled SRAM operation. Read-decoupled 7T and 8T cells improve stability by reducing the direct disturbance of internal storage nodes during access, but this benefit is usually accompanied by additional area or more involved access control [44–46]. Schmitt-trigger-based and latch-reinforced cells improve robustness and variation tolerance, particularly under subthreshold or near-threshold operation, yet these gains are typically achieved through increased transistor count and more complex internal structure [47, 49, 50, 57, 67]. Single-ended, TFET-based, asymmetric, and leakage-aware cells address power dissipation and reduced bitline activity, but often introduce asymmetry, technology dependence, or limitations in write flexibility and peripheral compatibility [51, 56, 60, 62, 63]. Similarly, disturb-free, half-select-resilient, and assist-enhanced cells improve writability and access reliability under low-voltage conditions, although these improvements are frequently associated with extra control signals, larger area, or added circuit overhead [52, 55, 59, 70].

The inclusion of SRAM array-level studies further clarifies that these trade-offs cannot be judged only at the isolated cell level. Several array-oriented works show that sensing speed, yield, leakage, voltage scaling, and reliability are strongly influenced by shared bitlines, common wordlines, peripheral circuitry, and statistical variation across large numbers of cells [2, 8, 72, 73]. In this context, a cell that appears favorable in terms of stability or low-voltage operation may still impose penalties at the array level through larger area, heavier bitline loading, added control complexity, or reduced density. Likewise, array-level approaches aimed at yield enhancement, voltage stacking, error tolerance, or in-array functionality demonstrate that architectural improvements are often obtained by accepting added peripheral overhead or more constrained design choices [74–77]. Thus, the broader understanding gained from combining bitcell and array perspectives is that SRAM design must be evaluated not only by local cell metrics such as SNM, write margin, or leakage, but also by how well a given cell structure scales within a practical array environment.

More recent secure and compute-oriented bitcells further show that SRAM cells are now expected to satisfy broader functional requirements beyond data storage alone, but this also increases the architectural burden placed on the cell itself [54, 64, 68, 71]. Taken together, the prior work indicates that SRAM bitcell research has made clear advances in stability, power efficiency, and application-specific robustness, while the inclusion of array-level studies reinforces

that these gains must ultimately be balanced against density, peripheral complexity, scalability, and implementation overhead. Therefore, the simultaneous achievement of strong read stability, reliable writability, low leakage, half-select resilience, and practical array compatibility remains a continuing design challenge.

1.4.5 DEVELOPMENT OF SRAM-IMC MACRO ARCHITECTURES

SRAM-based in-memory computing (SRAM-IMC) macros extend the role of SRAM from passive data storage to active in-memory processing by embedding selected computational capability within, or in close proximity to, the memory array. In the review literature, SRAM-CIM is most clearly understood across three interconnected levels: circuit, function, and application. At the circuit level, the macro is determined not only by the bitcell structure but also by peripheral support such as sensing paths, auxiliary logic, DAC/ADC interfaces, and timing control; at the function level, these circuit choices govern whether the macro supports digital Boolean operations, search, or mixed-signal accumulation; and at the application level, the resulting macros are targeted toward workloads such as CNN inference, classification, and related AI-edge processing [4]. More broadly, SRAM-CIM is attractive because it reduces repeated data transfer between processor and memory in data-centric workloads, while retaining the maturity, speed, and CMOS compatibility of SRAM as an embedded on-chip memory technology [1, 4].

A convenient way to introduce SRAM-IMC is through the standard 6T SRAM cell and array. In a conventional 6T array, data are stored in cross-coupled inverters and accessed through shared wordlines and bitlines. SRAM-IMC begins by observing that when multiple rows are activated, or when the bitline/sense path is intentionally interpreted differently, the array can perform more than storage alone. Depending on how the wordlines are driven, how the bitlines are sensed, and what peripheral circuitry is attached, the same 6T-based array can realize Boolean functions, search operations, or simplified inference primitives. This explains why many early SRAM-IMC macros were built around standard or lightly modified 6T SRAM structures: the memory fabric itself is reused as the physical medium for both storage and computation [4, 5, 78, 79].

From this baseline, SRAM-IMC macro development expanded in two major directions. One direction emphasized logic-oriented and search-oriented operation, where multi-row activation, modified sensing, or content-addressable access enabled in-memory Boolean functions such as AND, NOR, XOR, XNOR, IMP, and CAM-style comparison. The other direction targeted arithmetic-intensive inference workloads, especially binary, ternary, and multibit neural-network processing, where the macro design shifted toward product-sum evaluation, accumulation, and precision-scalable computation. This transition is reflected in X-SRAM, XNOR-SRAM, twin-8T CIM unit-macros, and voltage-mode or charge-domain SRAM-CIM macros, where the computation moved beyond functional reuse of storage arrays toward tightly integrated compute-memory datapaths involving peripheral conversion and accumulation circuits [6, 80–83]. The main review similarly distinguishes SRAM-CIM according to circuit-level realization and function-level capability, highlighting that the macro is shaped not only by the stored data format but also by the extent of peripheral intervention required to achieve the desired operation [4].

Before examining prior SRAM-IMC macros in detail, it is appropriate to clarify the meaning of throughput, energy, sensing, and scalability in a macro-level context. In a 6T SRAM-IMC macro, computation is not produced by a separate arithmetic unit external to memory; rather, it emerges from the interaction among stored cell states, wordline activation, shared bitline dynamics, and the associated peripheral circuitry. Consequently, the output of the macro is first manifested as a column-level electrical response and only thereafter interpreted by the sensing path. In this setting, *throughput* represents the effective computational output rate of the macro and is governed not only by operating frequency, but also by row activation parallelism, column-level output generation, and any additional cycles required for sensing, quantization, or accumulation. *Energy* denotes the total energy consumed in producing a valid computational result and therefore includes precharge activity, wordline switching, bitline development, sensing, reference generation, and peripheral conversion overhead rather than cell switching alone. *Sensing* refers to the reliable discrimination or quantization of the bitline response that carries the computational result, and it thus determines correctness, output separability, and attainable precision. Finally, *scalability* describes the extent to which the same computational principle remains effective as array dimensions, activated row count, operand precision, and peripheral complexity increase, since larger arrays generally intensify parasitic loading, latency, and variation sensitivity. Accordingly, in SRAM-IMC macros, these parameters are not independent figures of merit, but coupled system-level quantities that must be interpreted together in any meaningful performance assessment [4, 82].

More recent SRAM-IMC macros further broaden the design space through higher configurability, improved robustness, and support for wider application classes. Advanced-node macros now support multi-bit input, weight, and output handling at very high energy efficiency, while more recent works introduce variable-resolution activation, PVT-insensitive analog charge computation, and even floating-point segmented compute-in-memory organization [83–86]. In parallel, the broader SRAM review reminds us that CIM viability still depends on the underlying reliability of SRAM operation itself, since low-voltage behavior, variation, and peripheral complexity remain critical constraints even when computation is brought into memory [1]. Taken together, these trends show that SRAM-IMC macro design has evolved from standard-cell reuse and proof-of-concept logic-in-memory toward more configurable, precision-aware, and application-conscious macro architectures.

The progression outlined above is summarized in Table 1.4, which highlights how SRAM-IMC macros have evolved from 6T-compatible logic and search functions to more specialized mixed-signal, inference-oriented, and precision-configurable compute-memory architectures.

Table 1.4: Overview of SRAM-Based In-Memory Computing Architectures at the Bit-Cell and Macro Levels

Year	Reference	Technique	Observations
2016	Jeloka et al. [5]	Configurable 6T SRAM (TCAM/BCAM/SRAM)	• Push-rule based 6T operation • NOR/AND logic in memory • Multi-mode memory reuse • Minimal bitcell modification
2017	Dong et al. [87]	4+2T SRAM for Search & IMC	• BCAM/TCAM functionality • OR/AND/XOR operations • Area-efficient cell design • Near-threshold operation
2017	Zhang et al. [78]	6T SRAM-Based ML Classifier	• Dual SRAM/classify modes • In-memory ML inference • Comparable MAC accuracy • Reduced data movement
2018	Kang et al. [79]	6T In-Memory Inference Processor	• SVM, kNN, MF, TM support • End-to-end inference • Algorithm-aware memory use • Throughput improvement
2018	Agrawal et al. [6]	X-SRAM (8T Boolean IMC)	• NAND/NOR/XOR/IMP • Asymmetric SA usage • Logic-centric IMC focus • Extra transistor overhead
2020	Yin et al. [80]	XNOR-SRAM Macro	• Binary/ternary DNN MAC • 7T/8T pipelined macro • High throughput DNNs • Limited numerical precision
2020	Si et al. [81]	Twin-8T CIM Unit-Macro	• Multibit CNN MAC support • Two's-complement processing • Edge-AI optimized macro • Peripheral overhead
2020	Chiu et al. [88]	Configurable 6T CIM Macro	• 1–8 bit precision modes • Near-memory digital logic • Area-efficient scaling • Control complexity
2021	Sinangil et al. [84]	7-nm IMC SRAM Macro	• Multi-bit I/W/O support • Ultra-high TOPS/W • Advanced-node dependence • Peripheral-dominated area
2022	Mu et al. [82]	Voltage-Mode CIM SRAM	• Voltage-domain accumulation • Row-wise ADC • Mixed-signal trade-offs • Calibration requirement

Year	Reference	Technique	Observations
2023	Damodaran et al. [83]	SRAM IMC Macro with ΔΣ Activation	• Variable-resolution activation • Charge-domain compute • Improved linearity control • ADC complexity
2025	Chen et al. [85]	PVT-Robust Analog IMC SRAM	• In-situ multi-bit charge compute • 6T thin-cell compatible • PVT-insensitive operation • Analog calibration overhead
2025	Zhan et al. [86]	Floating-Point IMC Macro	• Segmented IMC architecture • Floating-point support • Broader workload coverage • Increased control complexity

As summarized in Table 1.4, SRAM-IMC macro research has developed through a broad sequence of architectural directions rather than along a single uniform path. Early macros largely focused on functional reuse of conventional SRAM structures, demonstrating that logic, search, and simple inference tasks could be mapped onto standard or lightly modified arrays with minimal disturbance to storage density [5, 78, 79, 87]. Later works increasingly targeted arithmetic-rich and neural-network-oriented workloads, where throughput, accumulation capability, and precision handling became central, and where peripheral circuits such as sense amplifiers, ADCs, DACs, and accumulators assumed a more dominant role in macro behavior [80–82, 84, 88]. More recent designs further indicate a shift toward reconfigurable, analog-robust, or numerically broader macros, including variable-resolution activation, PVT-insensitive analog computation, and floating-point compatible compute-in-memory structures [83, 85, 86]. Viewed together with the review literature, this progression suggests that SRAM-IMC macro design is no longer defined only by whether computation occurs inside memory, but by how effectively the macro balances bitcell compatibility, peripheral complexity, precision, energy efficiency, and application flexibility.

1.4.6 KEY OBSERVATIONS FROM PRIOR SRAM-IMC ARCHITECTURES

Although prior SRAM-IMC macro designs demonstrate that meaningful computation can be embedded within or near the SRAM array, the reported architectures generally optimize selected macro-level attributes rather than providing a uniformly balanced solution across all practical requirements. Macros based on standard or lightly modified 6T SRAM structures are attractive because they preserve compatibility with dense memory organization and enable logic, search, or simple inference operations with limited disturbance to the baseline array. However, their computational capability is often restricted unless additional sensing or peripheral support is introduced. In contrast, mixed-signal and arithmetic-oriented macros improve computational density and energy efficiency for neural-network style workloads, but these gains are frequently accompanied by stronger dependence on ADCs, DACs, reference generation, accumulation circuits, and analog signal integrity. As a result, improvements in throughput are often linked to

added quantization overhead, reduced sensing margin, or increased sensitivity to process, voltage, and temperature variations. Similarly, recent reconfigurable and precision-scalable macros broaden the functional scope of SRAM-IMC, yet they also show that greater flexibility usually requires higher control complexity and larger peripheral burden. Therefore, the literature suggests that the central challenge in SRAM-IMC macro design is no longer merely to demonstrate in-memory functionality, but to realize a macro that can jointly balance bitcell compatibility, sensing reliability, energy efficiency, computational throughput, and scalability under practical operating conditions.

1.4.7 ARRAY-LEVEL ORGANIZATION AND EVOLUTION OF SRAM-IMC ARCHITECTURES

SRAM-IMC array architectures extend the idea of in-memory computation beyond isolated bitcell or macro-level demonstrations by organizing large numbers of memory cells, shared wordlines, shared bitlines, and peripheral support circuits into a structured computational fabric. At this level, the central design question is no longer only whether a given in-memory operation can be realized, but how efficiently and reliably the same principle can be sustained across a larger array while preserving acceptable throughput, energy consumption, sensing margin, and implementation overhead. Consequently, SRAM-IMC array design must account not only for the computational mechanism inside the cell or macro, but also for array organization, dataflow, activation strategy, peripheral sharing, accumulation paths, and scaling effects introduced by larger dimensions and higher precision requirements.

Early SRAM-IMC array research largely explored how existing SRAM organizations could be reused as a computational substrate without completely abandoning the storage-oriented structure of conventional caches and on-chip memories. In this line of work, the array itself became the locus of computation, with bit-serial execution, in-cache mapping, and direct exploitation of shared wordline and bitline resources to support inference-oriented workloads. Representative examples include Neural Cache, SRAM-based CNN acceleration schemes, and array-level machine-learning classifiers, where the central objective was to reduce data movement by performing useful operations in place within large SRAM structures rather than transferring operands repeatedly to external arithmetic units [89–91]. These works established an important architectural foundation for SRAM-IMC arrays by showing that the usefulness of in-memory computing depends not only on the local bitcell or macro mechanism, but also on how data placement, cache organization, and array-level parallelism are orchestrated across a larger storage hierarchy.

As SRAM-IMC arrays matured, the research emphasis shifted from functional reuse alone toward higher computational density and richer arithmetic capability. This gave rise to array-level architectures supporting in-memory multiply–accumulate, column-parallel accumulation, charge-domain computation, and voltage-mode processing for neural-network acceleration [92–95]. In these designs, the array is no longer treated merely as a Boolean or search-capable memory fabric; instead, it becomes an analog or mixed-signal computational engine in which multiple

rows contribute simultaneously to a column response that encodes dot-product or accumulation behavior. Such an approach can improve throughput and reduce movement overhead, but it also makes the array increasingly dependent on peripheral circuits for reference generation, signal conditioning, quantization, and output accumulation. Consequently, the performance of these architectures is governed as much by sensing quality and peripheral efficiency as by the underlying storage-cell organization.

A further stage of development introduced transpose-array structures, bit-parallel precision scaling, and training-capable or reconfigurable array organizations. In this class of architectures, the array is explicitly designed to support not only inference but also broader dataflow patterns, including multibit operation, transpose-style mapping, and on-chip training support [96–98]. These works are important because they move the discussion from isolated in-memory computation primitives toward full array-level computational frameworks in which the organization of rows, columns, and peripheral blocks is matched to the target workload. At the same time, they show that improved computational flexibility is usually accompanied by stronger demands on control logic, reference management, sensing accuracy, and accumulation support, thereby increasing the architectural burden of the array.

More recent SRAM-IMC array designs increasingly emphasize scalability, programmable operation, and peripheral-aware optimization. Digital-friendly large-capacity arrays, charge-sharing ADC based arrays, sparsity-aware sensing strategies, reconfigurable 9T organizations, and adaptive digital reduction structures all indicate that the current state of the art is no longer concerned only with demonstrating in-array functionality, but with sustaining acceptable efficiency and correctness as array size, operand precision, and application diversity increase [99–105]. This progression makes clear that SRAM-IMC array design is fundamentally a co-optimization problem involving architecture, sensing, peripheral overhead, and workload mapping. The principal developments reported in this direction are summarized in Table 1.5.

Table 1.5: SRAM-IMC Array Designs Focusing on Architecture, Sensing, and Peripheral Overheads

Year	Reference	Technique	Observations
2018	Eckert et al. [89]	Neural Cache (In-Cache SRAM Compute)	• Bit-serial compute in cache SRAM • DNN operators mapped to cache • Reduces data movement • Cache layout and mapping constraints
2018	Abtahi et al. [90]	CNN Acceleration via SRAM Compute	• Convolution mapped to SRAM arrays • Direct and FFT-based flows • Algorithm–memory co-design • Bandwidth-aware execution
2018	Gonugondla et al. [91]	SRAM Array for In-Memory ML Classification	• Standard 6T SRAM array reuse • Array-level classification • Training/control overhead • Robustness via system tuning

Year	Reference	Technique	Observations
2019	Biswas and Chandrakasan [93]	CONV-SRAM Array for CNNs	• In-memory dot-product computation • Charge-domain compute in SRAM • Energy-efficient CNN acceleration • Analog/non-ideal sensing trade-off
2019	Kim et al. [94]	Mixed-Signal 6T SRAM IMC Array	• Voltage-mode accumulation • Row-by-row ADC operation • Mixed-signal array organization • Peripheral and ADC overhead
2020	Ali et al. [92]	Multi-Bit MAC in SRAM Array	• In-array dot-product computation • Column-parallel accumulation • Improves throughput • Calibration and swing limits
2020	Kang et al. [95]	SRAM-Based CIM Array for Neural Acceleration	• Mixed-signal MAC at array level • Neural-oriented array organization • Energy–accuracy trade-off • Peripheral-dominated scaling
2020	Wang et al. [106]	Compute SRAM for Programmable Vector Computing	• Bit-serial logic/arithmetic in SRAM • Programmable vector operations • Broad workload support • Mapping and peripheral overhead
2020	Lee et al. [96]	Bit-Parallel 6T SRAM IMC Array	• Reconfigurable bit-precision support • Bit-parallel in-array computation • Precision-scalable organization • Added control/peripheral complexity
2020	Jiang et al. [97]	CIMAT: Transpose SRAM Array for On-Chip Training	• Transpose-array organization • Supports on-chip training • Extends IMC beyond inference • Higher architectural complexity
2020	Su et al. [98]	Two-Way Transpose Multibit 6T SRAM CIM Array	• Inference–training array support • Two-way transpose organization • Multibit in-array operation • Peripheral and control overhead
2021	Su et al. [99]	Digital-Friendly SRAM-IMC Array	• 384 kb 6T-SRAM CIM macro • 8b precision AI-edge operation • Digital control emphasis • Control-logic overhead
2022	Lee et al. [100]	Low-Cost 7T SRAM-IMC Array	• Bitline charge-sharing ADCs • Parallel column readout • Reduced ADC cost • PVT sensitivity trade-off
2023	Zhang et al. [101]	TensorCache: SRAM-Based In-Cache IMC	• Tensor operations in cache SRAM • Compute near data • Targets bandwidth-bound ML • System-level integration cost

Year	Reference	Technique	Observations
2023	Xiao et al. [102]	SRAM IMC Array with Sparsity-Optimized ADC	• Hierarchical capacity attenuation • Input-sparsity-optimized ADC • 4b MAC-oriented array design • Analog/peripheral design overhead
2024	Li et al. [103]	Reconfigurable 9T SRAM-IMC Array	• Redundant unit and digital operation • Boolean logic and MAC support • Multi-mode array behavior • Control complexity and overhead
2024	Lee et al. [104]	P-8T SRAM IMC Array with In-SRAM DAC/ADC	• BL charge-sharing based DAC/ADC • High reported energy efficiency • Linear MAC-oriented array design • Mixed-signal peripheral dependence
2025	Park et al. [105]	Scalable Digital CIM Array with Adaptive Adder Tree	• Precision-controllable digital CIM • Adaptive adder-tree architecture • Flexible DNN inference support • System-level optimization focus

As summarized in Table 1.5, SRAM-IMC array research has progressed from early demonstrations of in-cache and in-array computation toward increasingly specialized architectures that emphasize arithmetic capability, sensing support, and scalable system integration. Initial works largely showed that conventional SRAM organizations could be repurposed for inference-oriented computation by exploiting in-place data access and array-level parallelism, thereby reducing external data movement and improving effective bandwidth utilization [89–91]. Subsequent designs expanded this concept toward mixed-signal and multibit array operation, where dot-product evaluation, column-parallel accumulation, and charge-domain or voltage-domain processing enabled higher computational density for neural-network workloads, but also increased dependence on calibration, sensing quality, and peripheral efficiency [92–95]. Later architectures introduced transpose-array organizations, bit-parallel precision scaling, inference–training support, and reconfigurable digital-friendly operation, indicating that SRAM-IMC arrays are no longer viewed merely as proof-of-concept compute fabrics, but as structured computational substrates whose usefulness depends on coordinated array organization, workload mapping, and peripheral design [96–99]. More recent works further show that scalability is increasingly governed by ADC overhead, sparsity-aware sensing, precision controllability, and adaptive reduction structures, so that the dominant design problem at the array level is the joint optimization of computational throughput, sensing reliability, peripheral overhead, and architectural flexibility rather than the realization of in-memory computation alone [100–105].

1.4.8 KEY OBSERVATIONS FROM PRIOR SRAM-IMC ARRAY ARCHITECTURES

Although the literature demonstrates that SRAM-IMC can be extended from isolated bitcell or macro concepts to larger computational arrays, the reported architectures still tend to optimize selected array-level objectives rather than providing a uniformly balanced solution across all practical requirements. Early in-cache and array-reuse approaches showed that meaningful

inference-oriented computation could be mapped onto existing SRAM organizations with reduced data movement and improved locality, but these benefits were often closely tied to specific cache layouts, operand-mapping strategies, or bit-serial execution styles, which limited generality and imposed nontrivial system-level constraints [89–91]. As the literature moved toward arithmetic-rich and neural-network-oriented arrays, the computational capability of SRAM-IMC improved significantly through column-parallel accumulation, charge-domain or voltage-domain evaluation, and multirow contribution to a shared column response. However, these gains also made the array increasingly dependent on calibration, reference generation, sensing quality, and peripheral conversion accuracy, so that performance improvement at the computational level was frequently accompanied by increased sensitivity to analog non-idealities, signal swing limitations, and process-voltage-temperature variations [92–95].

A second important observation is that scalability in SRAM-IMC arrays is not governed by array size alone, but by the interaction among array dimensions, simultaneously activated rows, precision requirements, and peripheral sharing. As the activated-row count increases, the available sensing margin may shrink because the computational result is increasingly encoded in a small differential or analog column response superimposed on substantial parasitic loading. Likewise, as array dimensions grow, bitline capacitance, latency, and column-to-column variation become more significant, making it difficult to preserve the same throughput and accuracy benefits that are often demonstrated in smaller or more controlled settings. In many reported designs, the array-level compute principle remains attractive, but the associated ADCs, DACs, accumulation trees, reference paths, or control logic begin to dominate area, power, and timing complexity. This is evident in transpose-array and training-capable architectures, where improved dataflow flexibility and multibit operation are achieved at the cost of stronger architectural overhead, more elaborate control, and tighter coordination between memory organization and workload mapping [96–99]. Similarly, more recent digital-friendly, sparsity-aware, and precision-controllable arrays improve particular aspects of scalability and programmability, yet they also confirm that peripheral-aware optimization is now inseparable from the core array design problem [100–105].

Thus, the broader lesson from the existing literature is that SRAM-IMC array design is no longer limited by the feasibility of in-memory computation itself, but by the difficulty of jointly balancing computational throughput, sensing reliability, precision support, architectural flexibility, and practical peripheral overhead in a scalable array environment. In other words, prior work has established several promising array-level directions, but a sufficiently balanced architecture that preserves efficient memory operation, supports useful compute functionality, and remains robust as array complexity increases is still a continuing design challenge.

1.5 RESEARCH GAPS

The literature reviewed in the preceding sections shows that substantial progress has been made in SRAM sense amplifiers, SRAM bitcell architectures, SRAM-based in-memory computing macros, and SRAM-IMC array organizations. However, the review also indicates that these problems have mostly been addressed in a fragmented manner, with most prior works improving

selected performance attributes while leaving important trade-offs unresolved. On the basis of the foregoing discussion, the following research gaps may be identified.

First, a clear gap exists in the design of SRAM sense amplifiers that can jointly provide fast sensing, robustness under reduced supply voltage, tolerance to process-induced mismatch, and practical implementation overhead. Prior sense amplifier designs have improved selected aspects such as speed, offset tolerance, or low-voltage read reliability, but these improvements are often accompanied by additional calibration circuitry, larger transistor count, stronger bias dependence, or increased control complexity. As a result, a balanced differential sensing architecture suitable for reliable SRAM read operation under variation-sensitive conditions remains an important open requirement.

Second, a gap remains in SRAM bitcell design with respect to the simultaneous achievement of read stability, write ability, low leakage, half-select resilience, and area efficiency. Although read-decoupled, Schmitt-trigger-based, single-ended, and assist-enhanced cells have significantly improved particular operating characteristics, these gains are usually obtained at the expense of larger transistor count, more control signals, reduced density, or stronger peripheral dependence. Thus, the literature still lacks a sufficiently balanced bitcell architecture that can sustain robust low-voltage operation while remaining practical for dense array implementation.

Third, in the area of SRAM-IMC macros, the literature reveals a gap between demonstration of in-memory functionality and realization of practically balanced compute-memory structures. Prior SRAM-IMC macros have successfully supported Boolean logic, search, classifier operation, and arithmetic-intensive neural-network primitives; however, many such designs either remain limited in computational flexibility or rely heavily on sensing circuits, ADCs, DACs, and auxiliary control paths. Consequently, a need remains for SRAM-IMC macro architectures that can improve computational utility and throughput without incurring excessive peripheral overhead or sacrificing robustness.

Fourth, a major gap persists at the SRAM-IMC array level, where scalability introduces additional constraints beyond those seen at the bitcell or macro level. Prior array-oriented works have demonstrated in-cache computation, mixed-signal accumulation, transpose-array organization, precision scaling, and reconfigurable operation, but these architectures often face growing challenges related to bitline loading, sensing margin degradation, peripheral dominance, latency increase, and variation sensitivity as array size and precision increase. Therefore, a scalable SRAM-IMC array architecture that can support useful in-memory computation while maintaining acceptable sensing complexity, reliability, and energy efficiency is still lacking.

More broadly, the literature indicates a clear gap in integrated co-design across the sensing, bitcell, macro, and array levels. Most prior works optimize one layer of the SRAM design hierarchy in isolation, whereas practical SRAM-IMC systems require coordinated interaction among reliable read sensing, robust cell behavior, efficient macro-level computation, and scalable array organization. Without such an integrated perspective, improvements achieved at one level may not translate effectively into practical throughput, energy, and reliability benefits at the system level.

These research gaps collectively motivate the present thesis. In particular, they point to the need for a coordinated design approach that addresses reliable SRAM sensing, robust low-voltage bitcell operation, efficient SRAM-IMC macro realization, and scalable dual-mode SRAM-IMC array architecture within a unified throughput-oriented framework.

1.6 PROBLEM STATEMENT

With the rapid growth of data-intensive and computation-oriented applications, the performance of modern digital systems is increasingly influenced by the speed and efficiency with which memory can support data storage and access. SRAM continues to be the most widely used on-chip memory because of its high speed and compatibility with standard CMOS technology. However, as technology scales and supply voltage is reduced, the reliable operation of SRAM becomes more difficult due to reduced sensing margin, process-induced variation, leakage, read/write conflict, and stability degradation. In addition, when SRAM is employed for in-memory computing, the design problem becomes more involved because the memory array is required to support not only storage but also computation through suitable sensing, peripheral support, and array organization.

The review presented in the previous sections shows that significant efforts have been made to improve different aspects of SRAM design. A number of sense amplifier structures have been reported to enhance sensing speed and low-voltage robustness. Similarly, several modified SRAM bitcells have been developed to improve read stability, write ability, leakage behavior, and half-select immunity. At the same time, SRAM-based in-memory computing macros and arrays have been proposed to reduce data-movement overhead and improve computational efficiency. However, most of these developments address the problem only from a particular design perspective. Improvement in one aspect is often associated with limitations in another, such as increased circuit overhead, added control complexity, larger area, stronger dependence on peripheral circuits, or reduced scalability at the array level.

Therefore, the problem considered in the present work is to develop an SRAM-based design framework for throughput enhancement in which the important requirements of sensing reliability, bitcell robustness, in-memory computation capability, and array-level scalability are addressed in a coordinated manner. In particular, there is a need to examine how an improved sense amplifier topology, a modified differential SRAM bitcell, an SRAM-based in-memory computation macro using an appropriate sensing strategy, and a reconfigurable SRAM-IMC array can be designed and integrated so that performance improvement is achieved without sacrificing practical energy efficiency, robustness, and implementation feasibility. The present thesis is undertaken in this direction.

1.7 RESEARCH OBJECTIVES

In view of the problem identified above, the present work has been undertaken with the following objectives:

1. Design and analysis of an improved sense amplifier topology for performance enhancement

in terms of variation tolerance.

2. Performance Enhancement of a modified differential bit cell for energy efficient SRAM.
3. Design and analysis of SRAM based in-memory computation (IMC) macro using modified sensing strategy.
4. Design and optimization of a reconfigurable SRAM-IMC array using modified SA and computation macro for throughput enhancement.

1.8 OBJECTIVE-WISE METHODOLOGY

The methodology adopted in the present work is objective-oriented. For each objective, the relevant literature has been studied, the design requirements have been identified, and the proposed circuit or architecture has been developed and evaluated through simulation and comparative analysis. The objective-wise methodology is described below.

METHODOLOGY FOR OBJECTIVE 1

The work related to Objective 1 begins with a detailed review of the sense amplifier architectures available in the literature in order to understand the main approaches used for fast sensing, low-voltage operation, and variation tolerance. From this study, the limitations of conventional and modified sense amplifier structures are identified with reference to read delay, power dissipation, sensing reliability, and sensitivity to parameter variations. An improved sense amplifier topology is then developed for differential SRAM read operation. The proposed architecture is analyzed under suitable SRAM read test conditions, and its performance is studied in terms of delay, power, and figure of merit. Its robustness is also examined under process and parameter variations. The obtained results are finally compared with those of representative existing sense amplifier designs to assess the extent of improvement.

METHODOLOGY FOR OBJECTIVE 2

The methodology for Objective 2 begins with the study of differential SRAM bitcell structures reported in the literature, particularly those intended for better read stability, write ability, and low-power operation. Based on this analysis, a modified differential SRAM bitcell is developed for energy-efficient SRAM operation. The proposed bitcell is then examined for its important static and dynamic characteristics, including stability and performance under different operating conditions. It is also considered together with the improved sense amplifier developed under Objective 1 so that cell operation and sensing behavior may be evaluated in a consistent manner. In addition to cell-level analysis, the bitcell is studied in an array environment to examine its suitability for practical SRAM realization. The design is evaluated in terms of yield, static noise margin, power, and reliability.

METHODOLOGY FOR OBJECTIVE 3

For Objective 3, the methodology starts with a review of SRAM-IMC macros reported in the literature so as to understand the computation styles, sensing schemes, and peripheral

requirements employed in existing designs. From this review, the important gaps in prior SRAM-IMC macro designs are identified, especially with respect to sensing strategy, throughput, power, and reliability. On this basis, an SRAM based in-memory computation macro is proposed using a modified sensing strategy. A suitable SRAM computational organization is selected and integrated with the modified sensing approach developed in the earlier objectives. The proposed macro is then simulated and analyzed for its computational behavior and performance. Its yield, throughput, power, and reliability are evaluated and compared with those of existing SRAM-IMC designs in order to establish the usefulness of the proposed architecture.

METHODOLOGY FOR OBJECTIVE 4

The methodology for Objective 4 begins with a review of reconfigurable SRAM-IMC architectures available in the literature, with emphasis on array-level organization, reconfigurability, and performance optimization. The limitations of the existing approaches are then identified. A reconfigurable SRAM-IMC array architecture is subsequently developed by combining the modified sense amplifier and the SRAM-IMC computation macro obtained from the earlier objectives. The proposed array is organized to support both memory mode and compute mode operation in a reconfigurable manner. Its functionality and performance are studied through simulation under both modes of operation. The array is finally evaluated with respect to throughput enhancement, and its overall behavior is analyzed from the viewpoint of practical array-level implementation.

1.9 THESIS ORGANIZATION

The present thesis has been organized in a sequential manner so that the work progresses from the general background and literature review to circuit-level development, macro-level realization, array-level integration, and finally to the conclusions drawn from the study. The arrangement of the chapters follows the logical flow of the proposed research objectives, beginning with the development of reliable read sensing, followed by bitcell design, in-memory computation at the macro level, and culminating in a reconfigurable SRAM-IMC array architecture. The organization of the thesis is as follows.

Chapter 1: Introduction and Literature Review

This chapter presents the background and motivation of the present work and discusses the growing importance of efficient on-chip memory design in view of the memory bottleneck and the increasing demand for data-intensive computation. It introduces the fundamentals of SRAM and SRAM-based in-memory computing, and presents a detailed review of the literature related to SRAM sense amplifiers, SRAM bitcells, SRAM-IMC macros, and SRAM-IMC arrays. The chapter further examines the major performance trends, trade-offs, and limitations reported in prior work. On the basis of this review, the research gaps are identified, the problem statement is formulated, and the objectives, methodology, and overall organization of the thesis are presented.

Chapter 2: A Process Variation Tolerant Modified Sense Amplifier for Differential SRAM

This chapter is devoted to the design and analysis of an improved sense amplifier topology for performance enhancement in terms of variation tolerance. It begins with the discussion of differential SRAM read operation and the role of sense amplifiers in determining delay, power dissipation, and reliability. The proposed sense amplifier structures are then introduced and their circuit operation is described in detail. The chapter presents simulation-based evaluation of the proposed designs with respect to delay, power, and power-delay product under different loading and operating conditions. Their behavior under process and parameter variations is also examined in order to assess robustness. The obtained results are compared with representative existing sense amplifier designs to establish the improvement achieved by the proposed approach.

Chapter 3: Proposed 8T SRAM Cell Design for Improved Performance

This chapter deals with the development of a modified differential SRAM bitcell intended for improved and energy-efficient SRAM operation. It begins with a discussion of the limitations of conventional and previously reported SRAM bitcells, especially with respect to read stability, write ability, half-select behavior, and power dissipation. The proposed 8T SRAM bitcell is then presented and its operation is explained in detail. The chapter includes the evaluation of the bitcell in terms of static noise margin, write margin, delay, dynamic power, and reliability under different supply voltages and operating conditions. The interaction of the bitcell with the improved sense amplifier is also considered so that cell and sensing performance may be understood in a consistent framework. In addition, the chapter examines the suitability of the proposed bitcell for practical array implementation.

Chapter 4: Differential 8T SRAM-IMC for Bitwise Logic Implementation

This chapter presents the design and analysis of an SRAM-based in-memory computation macro using a modified sensing strategy. It first discusses the basic concept of SRAM-IMC and the motivation for extending SRAM functionality beyond data storage to logic realization within the memory structure itself. The proposed differential 8T SRAM-IMC macro is then developed for bitwise logic implementation, and its operating principle is described in detail. The chapter explains how the modified sensing strategy is incorporated into the macro and how the computation is realized through the memory structure. The performance of the proposed macro is analyzed with respect to functionality, latency, throughput, power, and reliability. Comparative discussion with representative existing SRAM-IMC designs is also included in order to highlight the advantages and practical significance of the proposed scheme.

Chapter 5: Design and Analysis of a Dual-Mode Reconfigurable SRAM-IMC Array for Memory and Compute Operations

This chapter extends the work to the array level and presents the design and analysis of a dual-mode reconfigurable SRAM-IMC array capable of supporting both memory and compute

operations. The chapter begins with the motivation for array-level integration and discusses the challenges associated with scaling SRAM-IMC from macro-level demonstrations to a practical computational array. The proposed array architecture is then developed by integrating the modified sense amplifier, the proposed 8T SRAM bitcell, and the SRAM-IMC macro into a unified framework. Its organization and operation in memory mode as well as in compute mode are described in detail. The chapter further examines the behavior of the proposed array with respect to throughput, delay, power, energy, and scalability over different array sizes and operating voltages.

Chapter 6: Conclusion, Future Scope and Social Impact

This chapter summarizes the main findings obtained from the present investigation and highlights the important contributions made in the areas of sense amplifier design, SRAM bitcell development, SRAM-based in-memory computing macro realization, and reconfigurable SRAM-IMC array design. The major conclusions drawn from the analytical and simulation results are presented in a consolidated manner. The chapter also outlines the limitations of the present work and identifies possible directions for future research, particularly in relation to further optimization, larger array realizations, post-layout validation, and broader SRAM-IMC applications.

The thesis is followed by the list of references and the list of publications. The reference section includes the literature used to develop the thesis background, review prior SRAM sense amplifier, SRAM bitcell, SRAM-IMC designs, and support the comparative discussions across chapters. The list of publications summarizes the journal and conference papers arising from the doctoral work, reflecting the objective-wise outcomes and contributions of the present research.

CHAPTER – 2

A PROCESS VARIATION TOLERANT MODIFIED SENSE AMPLIFIER FOR DIFFERENTIAL SRAM

2.1 OVERVIEW OF THE CHAPTER

Static random-access memory (SRAM) arrays hinge on the ability of the sense amplifier (SA) to detect minute voltage differences on long, capacitively-loaded bit lines within a fraction of a clock cycle, making the SA the linchpin of every read operation. Modern SoC designs—whether high-performance AI accelerators or ultra-low-power IoT nodes—push the SA into increasingly aggressive voltage, noise-margin and variability corners, because any failure to resolve a '0' from a '1' propagates directly to system-level faults. As scaling exacerbates random threshold-voltage variations, advanced offset-tolerant topologies have emerged—redundancy schemes that vote across duplicated latches, and multi-sized transistors that bias device mismatch in favour of the desired switching polarity—both of which cut offset at the expense of increased static and dynamic power. Selecting the "right" SA therefore becomes a multidimensional optimisation: a voltage-mode design may suffice for moderate frequencies and wide supply rails, while deep-sleep or near-threshold memory blocks gravitate toward current-mode or charge-transfer solutions to regain performance headroom. Ultimately, based on the literature survey the SA dictates not only the achievable clock period of the cache or register file but also the lower voltage limit at which the entire SRAM sub-system can operate, making careful SA architecture choice as critical as the bit-cell topology itself in any memory-centric SoC. In response to the aforementioned requirement, we set forth the objective below:

"Design and analysis of an improved sense amplifier topology for performance enhancement in terms of variation tolerance"

Methodology.

1. Design of modified differential sense amplifier structures using varied topologies.
2. Compare the performance of the proposed sense amplifiers to identify the better SA design.
3. Analyse the selected SA under process variation and compare with other sense amplifier architectures.
4. Benchmark the performance of proposed sensing structure against pre-existing voltage and current SAs.

In this chapter, we propose two process tolerant SAs; a modified decoupled sense amplifier (MDSA) and a modified decoupled sense amplifier with NMOS foot-switch for performing differential-ended SRAM cell read operations. The Section 2.2 provides an introduction to sense

amplifiers for SRAM. The operation and structure of MDSA and MDSANF are discussed in Section 2.3 together with the impact of decoupling PMOS sizing on read performance as well as reliability of the SA under parameter variations. The Section 2.4 consists of performance analysis based on simulation results of the conventional voltage and current SAs with MDSA and MDSANF across the set of supply voltages. The chapter concludes with a summary of the key outcomes in Section 2.5.

2.2 INTRODUCTION

The backbone of numerous system-on-chip (SoC) devices used in crucial industries like the Internet of Things (IoT), biomedical implants, and wearable smart devices is a fast sensing, low-energy, and reliable Static Random Access Memory (SRAM). The SRAM is typically the dominant contributor to power consumption as well area occupied of the whole chip [107]. This puts the stability, delay, yield and energy consumption of SRAM at the forefront of the design strategy for high performance systems. As a result of the stringent design parameters for SRAM, the selection of bit cells, such as 6T, 7T [108–111], 8T [112], 9T [113] etc. together with peripheral circuitry including address decoder, column multiplexer [114], sense amplifier (SA) and write driver plays a crucial role in the design process. The ability to operate the system at lower voltages is one of the approaches employed in order to decrease the energy consumption thereby increasing the battery's life cycle. However, The operation of the SRAM at the minimum supply voltages with the minimum sized transistors for sub-45 nm technologies introduces various process, voltage and temperature (PVT) variations which when compared with operation at nominal supply voltages and larger device sizing, could result in a two-fold speed degradation [115].

The Sense Amplifier (SA) plays a significant role in the read operation of SRAM. The SA performs the amplification of a small differential signal present at the bit lines to larger rail-to-rail swing although single-ended SA architectures have also been adopted in SRAM [116, 117]. Since, SRAM cells are highly sensitive to process variations, the SA is also affected by the random variations caused by Random Dopant Fluctuations (RDF). The mismatch in threshold voltages has a vital impact in deteriorating the SRAM read yield and the access delay.

The SA topologies determine the performance and operation limits of the SRAM. The widely adopted voltage-mode SAs are based on cross-coupled inverter latches, which are simple and robust [118–121]. SAs have a small delay due to the positive feedback of cross-coupled inverters, while voltage-mode SAs require a significant voltage difference between a pair of bit lines due to their offset. Large delays result from the small bit cell driving large capacitive bit lines. Bit lines can be reduced in capacitance by using divided bit lines and two-level sensing [119, 121]. Since the local bit line connects fewer bit cells, there is a lower bit line discharge delay, but the bit line full voltage swing consumes more energy.

Current-mode SAs may have more advantages because of no extra intermediate voltage [122–126]. The current conveyor widely utilized in current-mode SAs uses four PMOS transistors to create a virtual short circuit through the bit lines as it reduces the sensing delay and is not

immune to bit line capacitance variations [127]. Also, a clamped bit line current-mode SA with low impedance was presented [122], which maintains a constant potential on the bit lines. Following this, most research on current-mode SAs has focused on circuit improvement based on the current conveyor and the clamped bit line technique to improve speed and energy efficiency.

The process variation has a large effect on SA offset with CMOS technology node scaling. SA offset caused by inherent mismatch at lower technology nodes has a negative impact on SRAM yield. Many solutions have been reported to reduce offset [128–133] with one of the most basic approaches being the use of enlarged transistors [129]. Multiple redundant SAs of same [134] and different sizes [135] were also used together with advanced calibration schemes to counter offset. These calibration-based SA topologies allow for high speed operation with lower power consumption while having a larger area overhead and increased complexity.

2.3 PROPOSED VARIATION TOLERANT MODIFIED DECOUPLED SENSE AMPLIFIER

Figure 2.1 (a) and 2.1 (b) illustrates the detailed circuit schematic of the proposed modified decoupled sense amplifier (MDSA) and MDSA with NMOS footswitch (MDSANF) respectively. Both the SA architectures are designed using the conventional inverter latch decoupled by the PMOS switches M6 and M7. The operation of the SAs can be divided into three segments. The three segments are the pre-charge, amplification and output propagation. The pre-charge segment requires the bit lines to be charged to supply voltage V_{DD} using M1-M3 PMOS transistors. The timing control signal (TCS) used here is PRE. The amplification stage is initiated by turning on decoupling switches (M6 and M7) by turning the TCS SA_EN off. The cross-coupled latch is comprised of two inverters formed by the M5, M8 and M5, M9 transistors respectively. The latch then acts as a high-gain positive feedback amplifier wherein a differential current generated across the drains of the M8 and M9 gives rise to large differential voltage across SAO+ and SAO-. The SAO+ is pulled high while the SAO- is pulled to GND for a read ”1” operation and vice versa. The outputs nodes of the SA are passed through inverters to attain better isolation and the maximum output signal swing. This contributes to lower delay during the read operation and improves the performance of the sense amplifier. The additional NMOS footswitch M10 used is similar to the ones used in conventional voltage-mode SAs. It serves as a sleep transistor to reduce leakage in the power gating aspect of the SA. The Figure 2.1 (c) and 2.1 (d) depict the transient response for MDSA and MDSANF respectively at V_{DD} = 0.8 V with C_{BL}(bit line capacitance) being 50fF. The transient response is used to derive the delay performance of the SA. The delay of the sense amplifier assumed to be the time between the 50% point of the SRAM word line (WL) and the 50% point of the differential voltage at SAO nodes [136].

2.3.1 IMPACT OF SIZING OF DECOUPLING PMOS ON SA READ PERFORMANCE

To decouple the cross-coupled inverters in the latch before the sense amplifier is enabled, the MDSA and MDSANF topologies employ the PMOS transistors M6 and M7. Figure 2.2 illustrates how widening these devices affects read behaviour over the supply range V_{DD} = 0.5–1.0 V. In

Figure 2.2 (a) and Figure 2.2 (d), the read delay decreases monotonically with width because a wider PMOS supplies more drive current than the extra gate and diffusion capacitance it introduces. The benefit is most pronounced above 0.8 V; at 1.0 V the 88 nm device trims delay by roughly one-third relative to the minimum-sized 44 nm case, whereas at 0.5 V mobility degradation limits the gain and delay remains above 1 ns for every width.

Power consumption, plotted in Figure 2.2 (b) and Figure 2.2 (e), is nearly constant across the width sweep except at 1.0 V, where the larger dynamic load $C_L V_{DD}^2 f$ of the widest devices becomes visible. Consequently, the power–delay-product curves in Figure 2.2 (c) and Figure 2.2 (f) show a clear minimum at the intermediate width of 88 nm; this point provides the best energy–speed compromise while maintaining safe read margins.

Comparing the two architectures, MDSANF consistently incurs a modest delay penalty of about 10 % at 0.6 V but achieves a small power advantage, reflecting the speed–energy trade-off introduced by its negative-feedback path. Because M6 and M7 participate only in the read phase

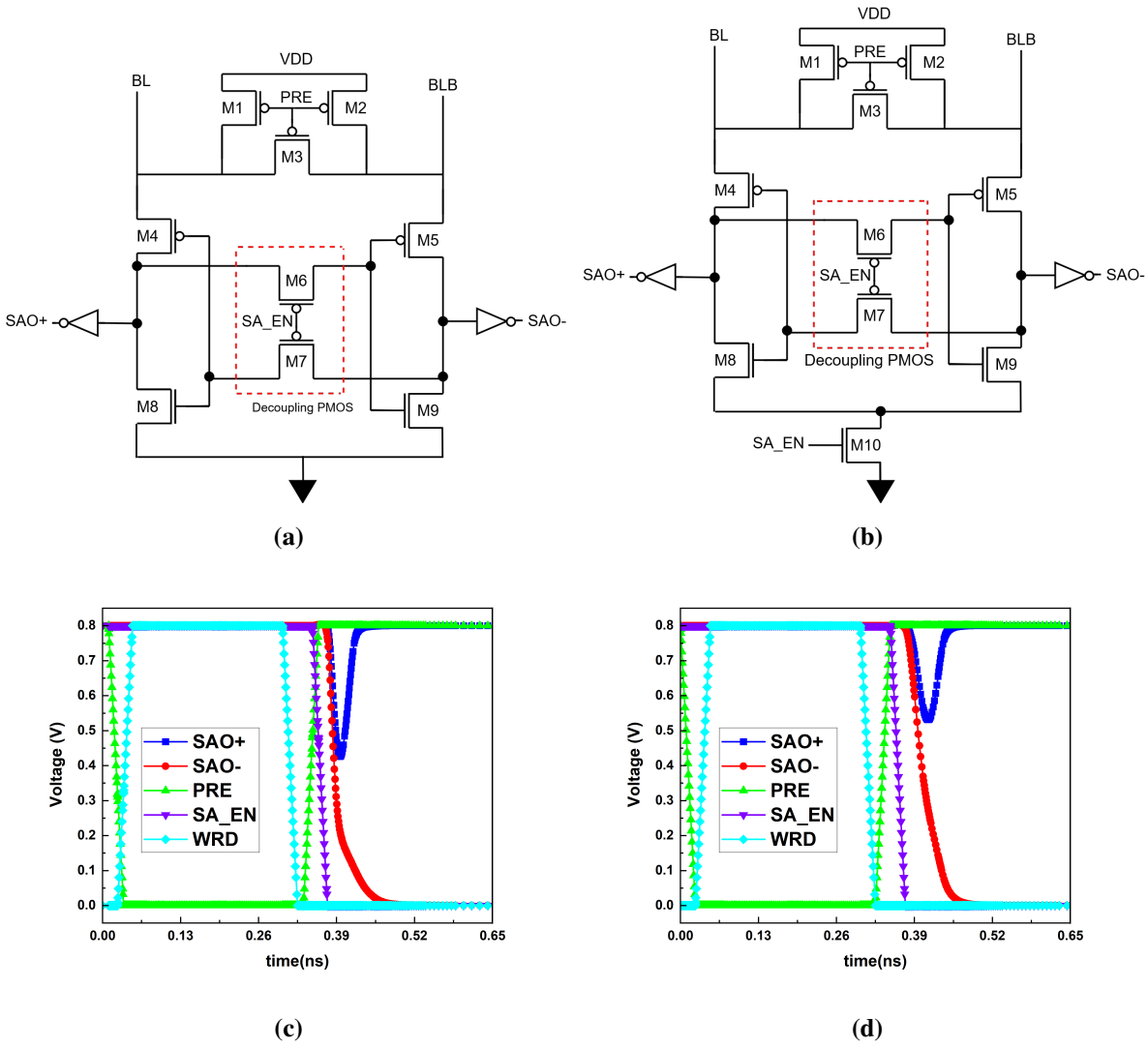


Figure 2.1: Circuit schematics for (a) MDSA and (b) MDSANF, along with the transient response showing timing control signals at V_{DD} = 0.8 V and C_{BL} = 50 fF for (c) MDSA and (d) MDSANF

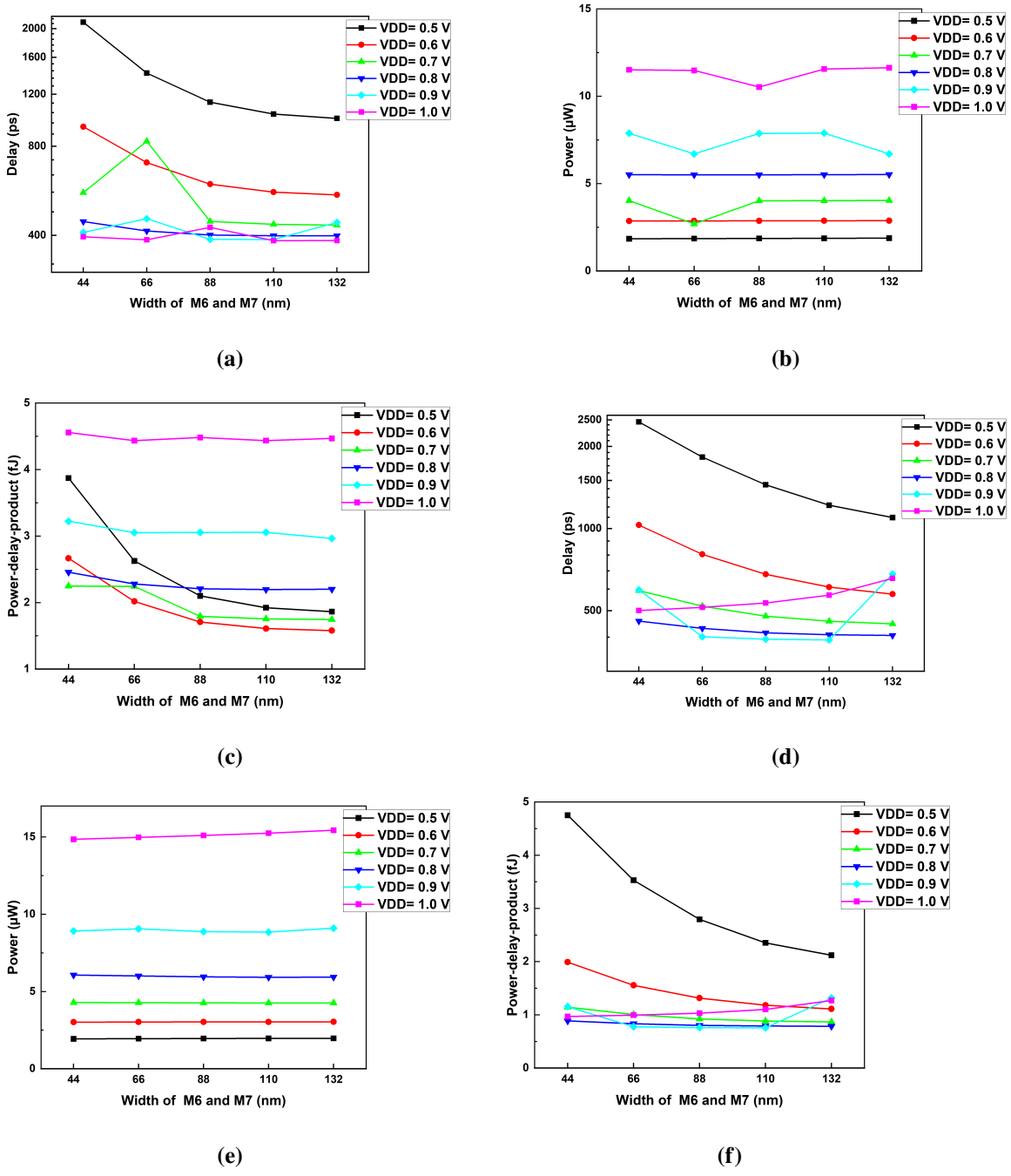


Figure 2.2: Investigating the impact of PMOS devices M6 and M7 width on MDSA and MDSANF performance: (a) Delay, (b) Power consumption, and (c) Power-delay-product for MDSA and (d) Delay, (e) Power consumption, and (f) Power-delay-product of MDSANF examined for different supply voltages

(see Figure 2.2), upsizing them does not disturb the write path, and the extra silicon area is confined to the sense-amplifier column. A width of 88 nm is therefore adopted for all subsequent best-case calculations: it minimises delay, reduces the power–delay product and, thanks to its larger channel area, also improves process-variation tolerance without violating the column-pitch constraints of the SRAM layout.

For SRAM cells expected to operate below 0.8 V, sizing M6 and M7 to 88 nm offers the

optimal balance of energy and speed; in high-performance paths fixed at 1.0 V a narrower 66 nm device may be selected to save dynamic energy with only a marginal delay penalty.

2.3.2 PARAMETER VARIATION TOLERANCE OF PROPOSED SA

RDF refers to the variations in the number of dopant atoms present in the transistor channel. As technology nodes decrease, the number of dopant atoms decreases significantly, reaching around ten atoms at 16 nm compared to a few tens at 45 nm [137].

This reduction in dopant atoms contributes to the inconsistency in transistor parameters, thereby increasing the variability in CMOS devices at smaller technology nodes.The effect of process variation on the MDSA and MDSANF is studied using Monte Carlo simulations for 10k data points. A read failure might happen as a result of V_{TH} fluctuations, which is defined as the read output for a read "1" operation going to zero rather than V_{DD}. The transient responses at V_{DD} = 0.8 V and C_{BL} = 50 fF for MDSA and MDSANF are depicted in Figure 2.3 (a) and 2.3 (b). The transient response depicts that there are zero read failures that take place during the SRAM read.

CMOS parameter variations can be classified into two main categories: front-end variations and back-end variations. Among these, front-end random variations are primarily influenced by the fluctuation in threshold voltage (VTH). One of the key contributors to front-end random variations is random dopant fluctuations (RDF). It is assumed that the threshold voltage fluctuation follows a Gaussian distribution with 3σ variation around the mean μ. The lowest read failures are used to determine a SA's yield.

A successful read operation in SRAM critically dependent on the symmetry of the majority of differential-ended SAs. The distribution of read delay for MDSA is shown in Figure 2.3 (c) for near threshold operation and Figure 2.3 (c) (inset) for other supply voltages. Similarly, the delay distribution for MDSANF depicted in Figure 2.3 (d) and 2.3 (d) (inset). The delay distributions are calculated from the MC simulations. The near threshold voltage operation shows deviation larger than a few hundred ps but goes down to tens of ps for higher supply voltages as is evident from Figure 2.3 (c) and 2.3 (d) for both MDSA and MDSANF.

To perform a performance analysis of a sense amplifier in an SRAM array, various design parameters can be considered, including the bit line capacitance. The bit line capacitance is a crucial parameter that affects the speed and overall performance of the sense amplifier. By considering different values of bit line capacitance, representing varying array sizes, a comprehensive analysis can be conducted to assess the impact on the delay, power consumption, and noise margin of the sense amplifier in the SRAM array. Other design parameters that can be considered for performance analysis include the array size (number of memory cells), bit line capacitance, supply voltage, and process technology. Each of these parameters influences the behaviour of the sense amplifier and can be used to evaluate its performance under different operating conditions and array configurations. A singular SA is used to perform read operation for a column of bit cells which can range upwards of 1024 in large SRAM array structures. The delay variation as a function of C_{BL} is shown in Figure 2.4 (a). The increase in C_{BL} causes larger

read delay due the SA being forced to discharge large capacitance to achieve a correct read. The delay shows a 4.77 times increase for MDSA for the bit line variation from 10 fF to 3000 fF and 6.25 times increase for MDSANF. To further justify the SA's robustness, the delay analysis for MDSA and MDSANF is extended across various global corners at V_{DD} = 0.8 V and C_{BL} = 50 fF

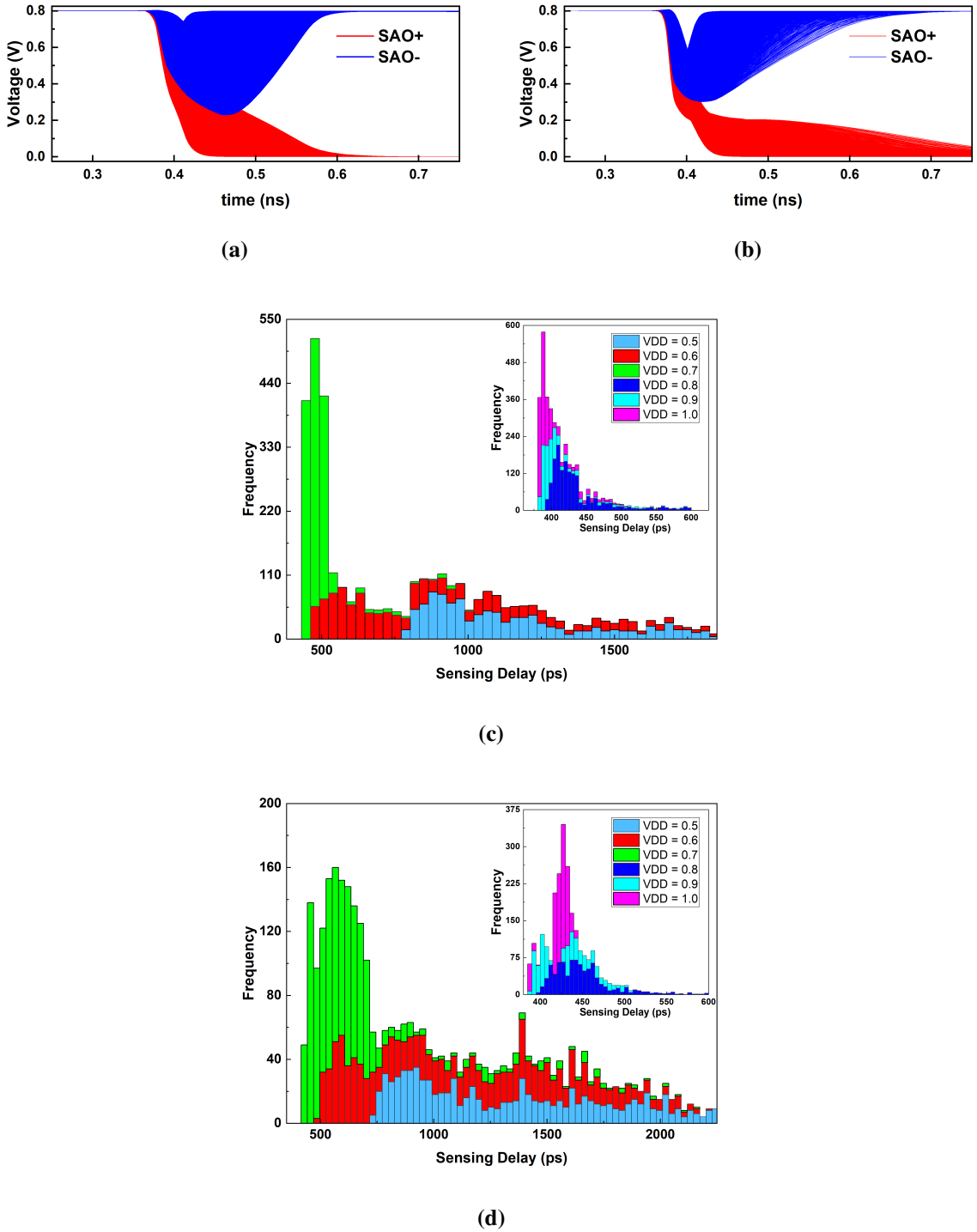
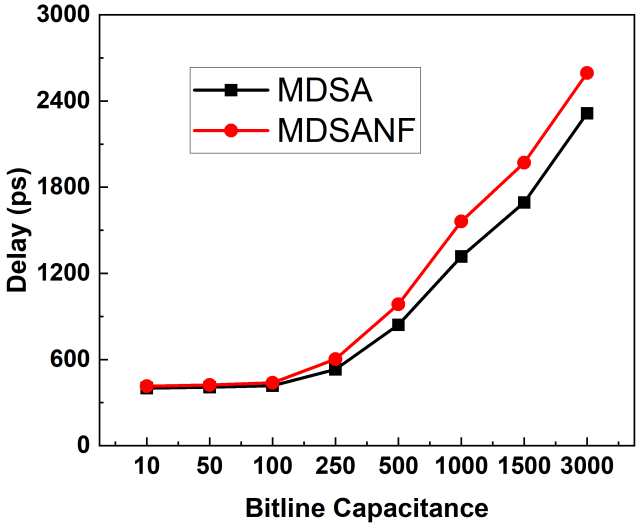
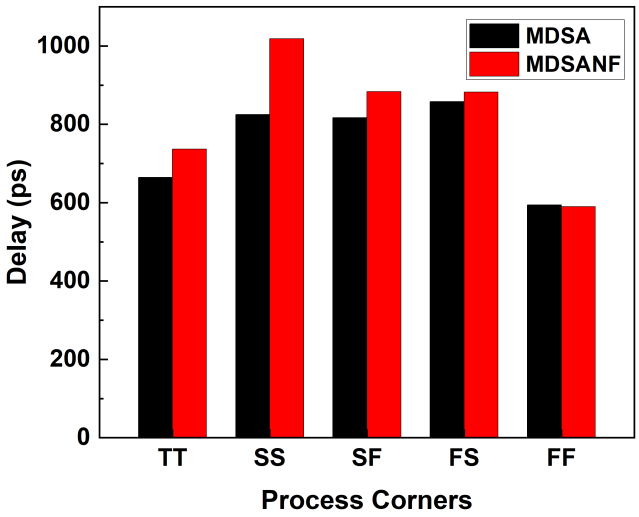


Figure 2.3: Transient response (10k Monte Carlo simulation) of (a) MDSA, and (b) MDSANF and calculated delay distribution at C_{BL} = 50fF for near threshold operation of (c) MDSA and (d) MDSANF (Inset: delay distribution at above threshold operation)



(a)



(b)

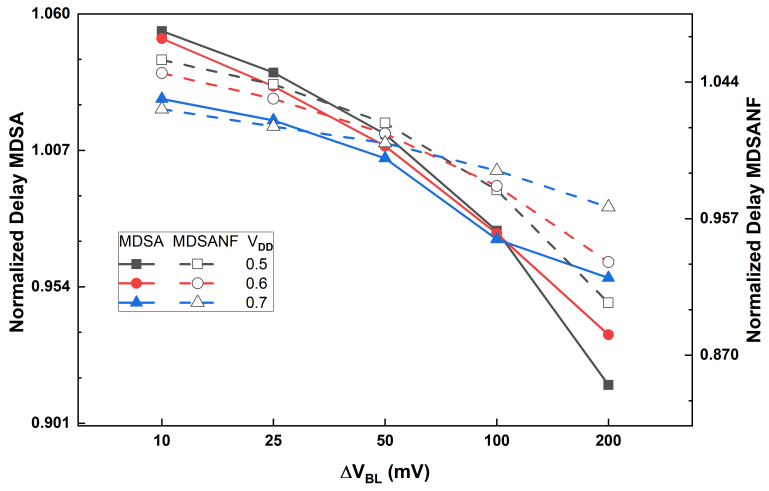
Figure 2.4: Delay variation as a function of (a) Bitline Capacitance and (b) Process corners for MDSA/MDSANF at VDD = 0.8 V

and these results are shown in Figure 2.4 (b).

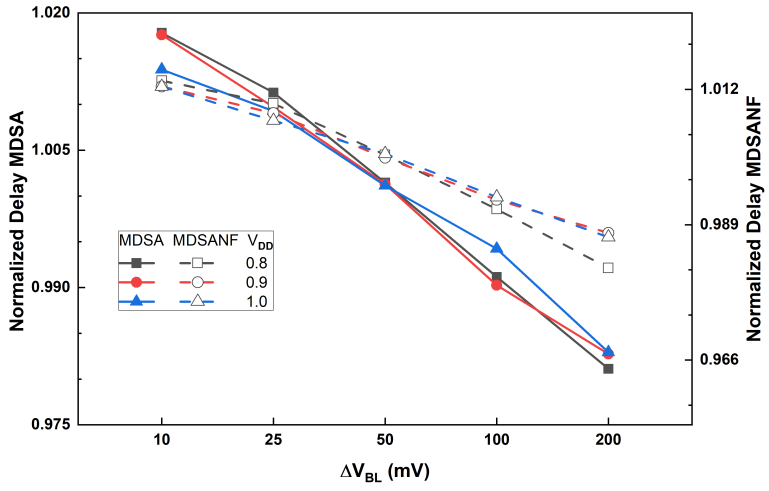
The normalized delay of the SAs as function of the input differential bit line voltage is shown in Figure 2.5 for the set of supply voltages. The normalized delay values show most deviation for near threshold operation of 0.5 V with deviation of 185.358 ps from fastest to slowest read delay and least deviation of only 10.932 ps at 1 V supply for MDSANF. Similarly, the largest delay deviation is 118.21 ps at 0.5 V while the smallest is 6.557 ps at 1 V for MDSA.

2.4 PERFORMANCE COMPARISON OF MDSA/MDSANF WITH EXISTING SA ARCHITECTURES

A meaningful evaluation of the proposed MDSA and MDSANF requires comparison with established SRAM sense-amplifier families under identical operating conditions. Since each topology senses and regenerates the bit-line signal through a different mechanism, the resulting read delay, power consumption, leakage, power-delay product, and area are governed by different circuit-level trade-offs. The following subsection classifies representative SRAM sense amplifiers according to their sensing principle and design objective, thereby providing the basis for the subsequent performance comparison of the proposed MDSA/MDSANF with existing voltage-mode and current-mode architectures.



(a)



(b)

Figure 2.5: Normalized Delay as a function of differential bit line voltage input for MDSA and MDSANF at (a) near threshold and (b) above threshold supply voltages

2.4.1 CLASSIFICATION OF SRAM SENSE AMPLIFIERS

SRAM sense amplifiers can be broadly classified according to the mechanism used to detect and regenerate the read signal from the bit-lines. At a high level, prior designs may be grouped into voltage-mode latch-based sense amplifiers, current-mode or hybrid sense amplifiers, and more advanced assisted or variation-tolerant architectures developed to improve robustness under low-voltage and scaled-technology operation. This classification is useful because each group reflects a different design emphasis, such as simplicity and compactness, faster sensing through reduced bit-line dependence, or improved tolerance to offset, mismatch, and process variation.

Voltage-mode sense amplifiers constitute one of the most widely adopted classes of SRAM read circuits because of their compact structure and strong regenerative amplification based on cross-

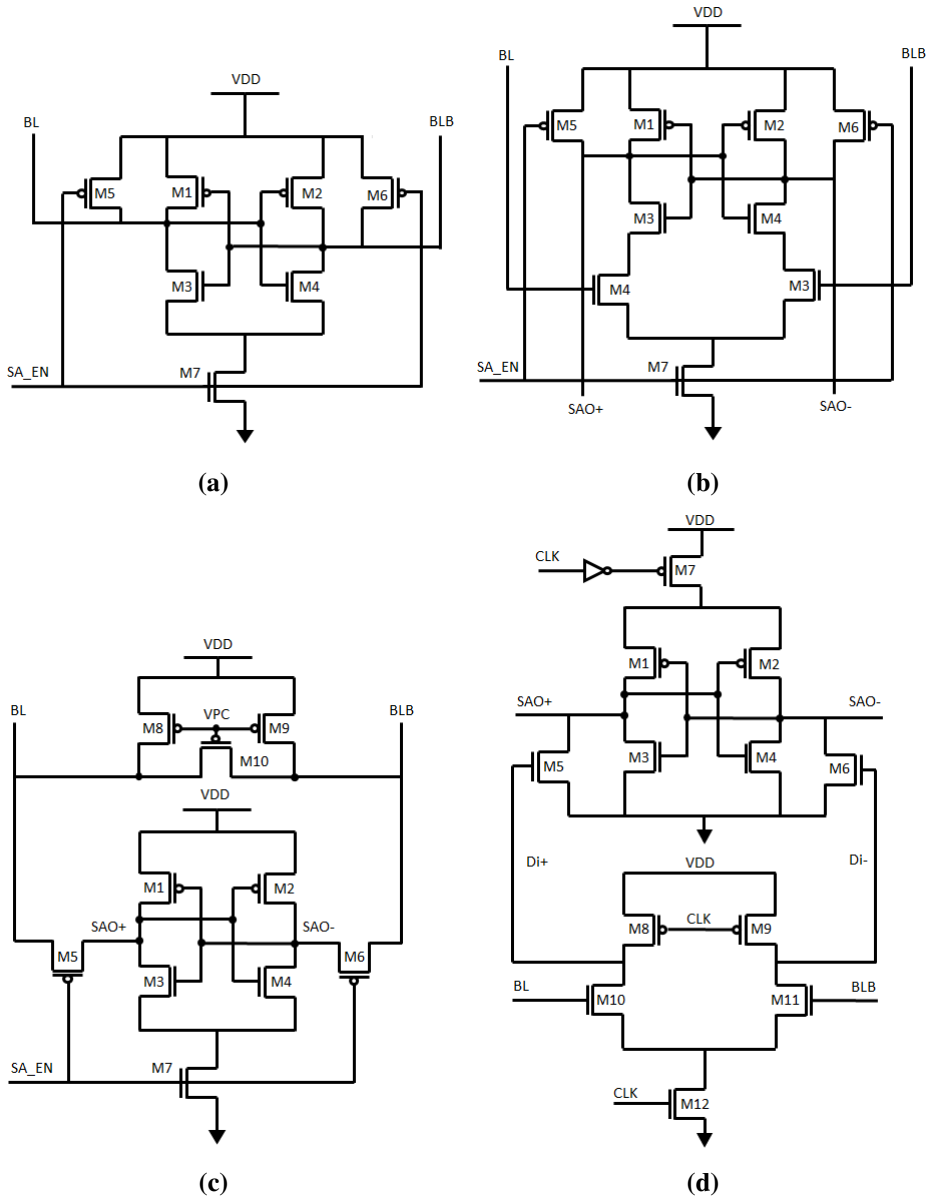


Figure 2.6: Circuit schematics for Voltage SAs (a) Conventional latch-type SA, (b) Current-controlled latch-type SA, (c) Decoupled latch-type SA, and (d) Double tail SA.

coupled inverter latches. In these architectures, sensing begins only after a sufficient differential voltage has developed on the complementary bit-lines, following which positive feedback rapidly regenerates the small input difference into a full-swing digital output. Representative voltage-mode architectures considered in this review are shown in Fig. 2.6, including the conventional latch-type sense amplifier, current-controlled latch-type sense amplifier, decoupled latch-type sense amplifier, and double-tail sense amplifier [42, 138–140].

The conventional latch-type sense amplifier shown in Fig. 2.6 (a) is one of the earliest CMOS-based sensing structures used in memory design. It consists of a cross-coupled inverter pair, and the regenerative action is enabled through the sense-enable signal, allowing a small bit-line voltage difference to be amplified to the output. Although this structure is simple and robust, its operation is affected by the large capacitance seen at the bit-lines, which leads to increased delay and power consumption. To improve upon this behavior, the current-controlled latch-type sense amplifier in Fig. 2.6 (b) introduces additional transistors that regulate the current flowing through the latch during sensing. This creates a high-input-impedance condition at the bit-lines and helps achieve faster sensing compared with the basic latch structure, particularly when the bit-line loading is significant [138].

Further improvements in voltage-mode sensing were achieved through input decoupling and staged regeneration. In the decoupled latch-type sense amplifier shown in Fig. 2.6 (c), bidirectional pass transistors are inserted between the bit-lines and the latch so that the cross-coupled inverters do not directly drive the large capacitive bit-lines during sensing. This decoupling reduces power consumption and improves sensing speed, provided that the bit-lines are properly precharged before the read operation [139]. The double-tail sense amplifier in Fig. 2.6 (d) extends this idea by separating the sensing path into two tail branches, one associated with the input bit-lines and the other with the latch regeneration stage. During operation, an input-dependent voltage difference is first developed at internal nodes and is then transferred to the cross-coupled latch for amplification. This architecture offers better isolation between input and output nodes and is particularly attractive for lower-voltage operation, although it requires a higher transistor count and more elaborate timing control [140].

Current-mode and hybrid sense amplifiers were introduced to reduce the dependence of SRAM read sensing on the development of a large differential bit-line voltage prior to regeneration. In these architectures, the sensing action is influenced more strongly by the current response of the bit-lines, often through clamped bit-line operation or current-conveyor based front-end structures, so that faster read decisions can be achieved before a large voltage swing appears on the bit-line pair. Representative current-mode and hybrid architectures considered in this review are shown in Fig. 2.7, including the clamped bit-line sense amplifier, hybrid current sense amplifier, modified clamped sense amplifier, and modified hybrid sense amplifier [141–144].

The clamped bit-line sense amplifier shown in Fig. 2.7 (a) employs dedicated clamp transistors to maintain the bit-lines near a nearly constant potential during sensing. This reduces the bit-line voltage excursion and enables the read decision to be governed by the differential current rather than by a large developed voltage difference. The hybrid current sense amplifier in Fig. 2.7 (b)

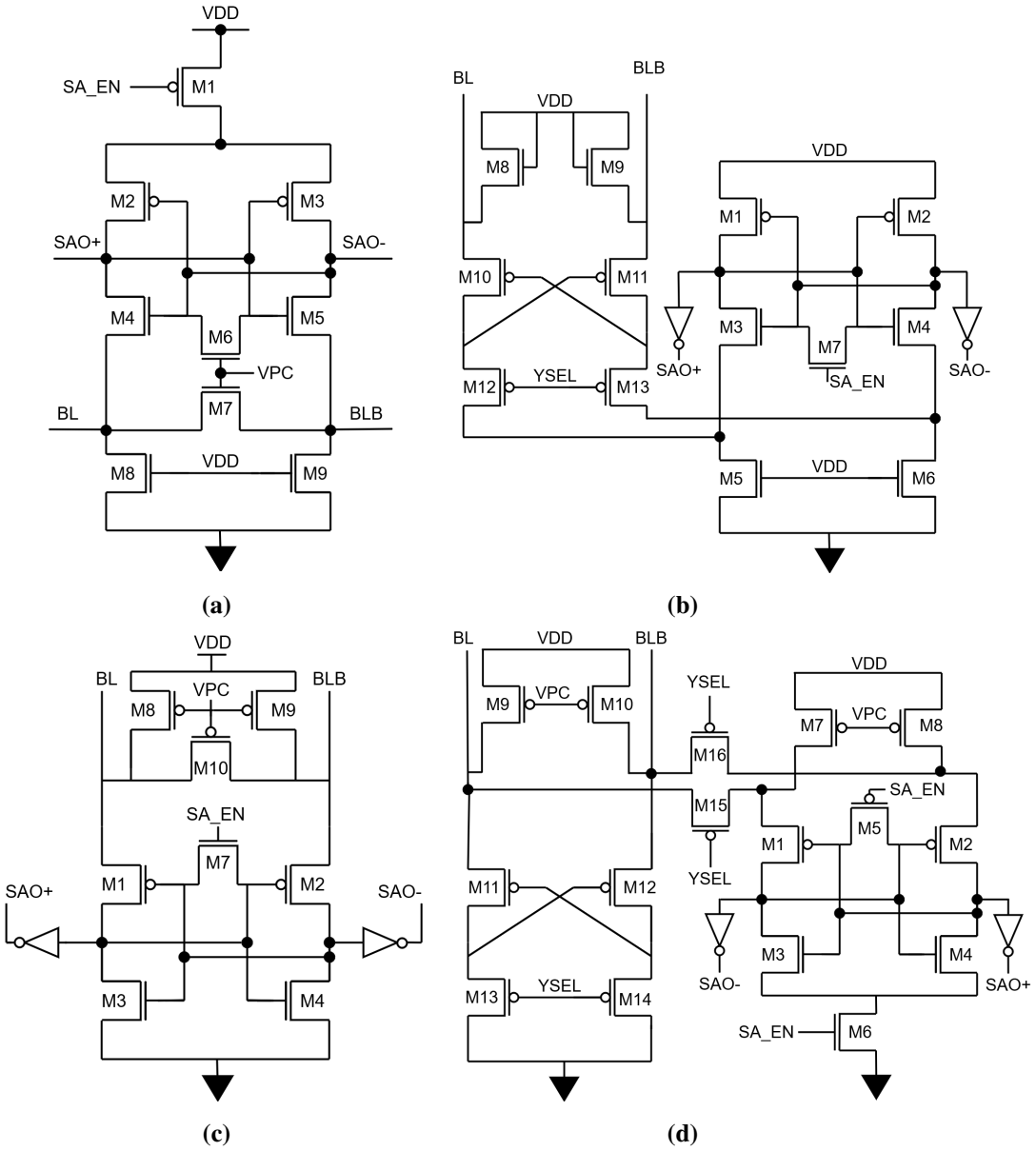


Figure 2.7: Representative current-mode and hybrid SRAM sense amplifier architectures: (a) clamped bit-line sense amplifier, (b) hybrid current sense amplifier, (c) modified clamped sense amplifier, and (d) modified hybrid sense amplifier.

extends this idea by incorporating a current conveyor formed by cross-coupled PMOS devices, which transports the bit-line current response to the latch stage without introducing direct voltage gain at the input. As a result, these architectures can achieve faster sensing and reduced sensitivity to large bit-line capacitance, although they require additional circuitry and biasing support compared with basic voltage-mode latch structures [141, 142].

Subsequent current-mode designs introduced further modifications to improve symmetry, tracking, and sensing reliability. In the modified clamped sense amplifier shown in Fig. 2.7 (c), an additional equalization device is incorporated together with a precharge stage so that the bit-lines begin the sensing event from a more balanced condition, allowing the small quiescent

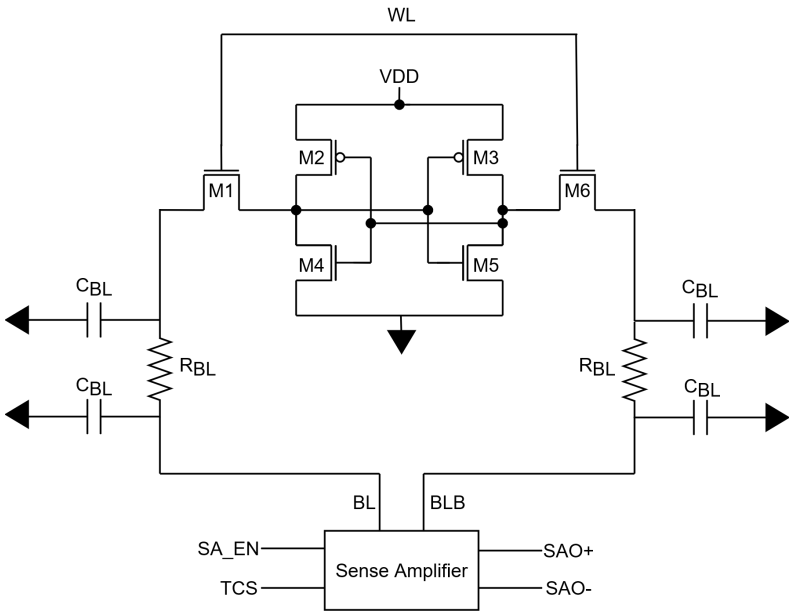


Figure 2.8: Simulation test-bench used for comparison of sense amplifiers.

differential voltage to be tracked and amplified more effectively [143]. The modified hybrid sense amplifier in Fig. 2.7 (d) combines the conventional current-conveyor concept with additional data-line coupling devices, so that the bit-line response can be transferred and regenerated with improved sensitivity under low-voltage operation [144]. Although these schemes improve speed and sensing robustness, they also increase transistor count, control complexity, and in some cases power consumption due to the added precharge, equalization, and coupling circuitry.

The comparative analysis of all the conventional SA architectures is presented in this section with all the SAs being designed using Predictive Technology Model [145] at 22 nm technology node. All the various SAs were designed using consistent sized transistors in order to conduct an objective analysis at various supply voltages. The conventional voltage and current-mode SAs have transistors with W/L = 44 nm/ 22 nm while all access transistors are sized with W/L = 1000 nm/ 22 nm. The read delay, power consumption and leakage have been calculated based on the simulation results. An important performance statistic for SA is the power-delay-product, or PDP for short. A stronger overall performance is shown by the lower PDP numbers. The identical word line (WL) and sense amplifier enable (SA_EN) timing signals are utilized while performing all simulations, together with the precise timing and control signals (TCS) as specified in the relevant literature. The simulation testbench schematic used to compare the sense amplifiers is shown in Figure 2.8. A single 6T SRAM cell is used together with a π-3 model for the bitline capacitance of large SRAM array structures. This is similar to test circuit in [146].

2.4.2 READ DELAY

The conventional SA architectures are simulated and the calculated read delay is shown in Figure 2.9. As is clearly evident, MDSA and MDSANF have the fastest operation for all supply voltages except the near threshold voltage operation at V_{DD} = 0.5 V.

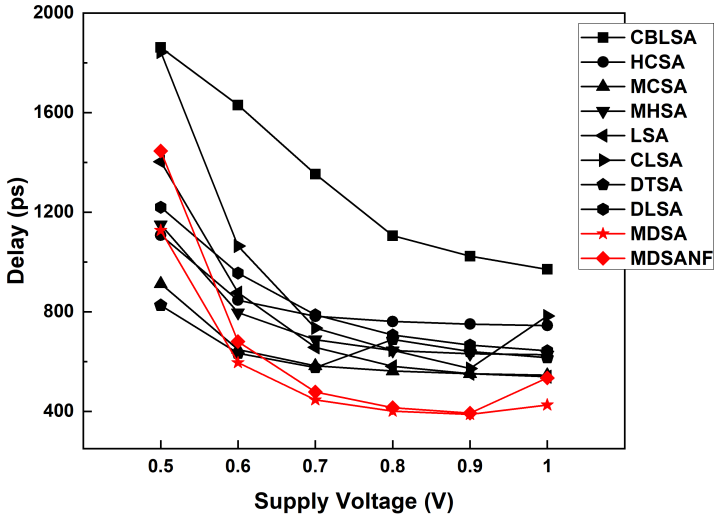
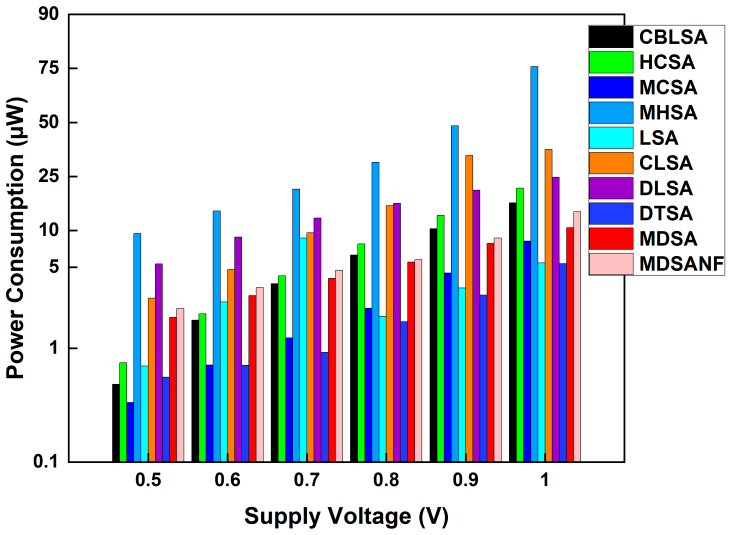


Figure 2.9: Delay comparison of all sense amplifiers as a function of the supply voltage at C_{BL} = 50 fF.

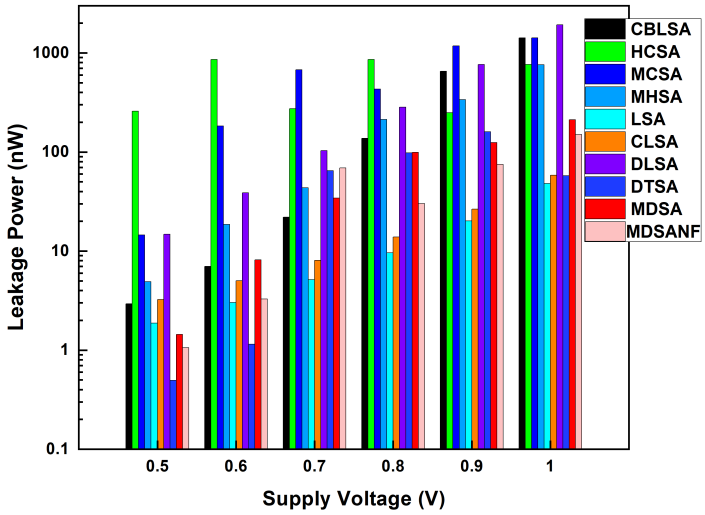
The proposed MDSA shows a 28.6% read delay improvement over MCSA while MDSANF shows a 26.13% improvement at V_{DD} = 0.8 V. The MCSA is the fastest among the voltage-mode and current-mode SAs followed by LSA and MHSA. MHSA, MCSA and DTSA exhibit the most consistent delay across voltage supply variation with DTSA having its best and worst delay 210.5 ps apart. The fastest operation at lowest applied supply is given by DTSA with MCSA being the second fastest as a result 10.49% increase in delay. Amongst the current SAs compared, only CBLSA has delay consistently larger than 1 ns across all the supply voltages applied. The conventional SA designs compared above are all simulated for a bit line capacitance modeled by 50 fF with ΔV_{BL} assumed to be 50 mV with transistors having no mismatch operated at TT/27°C. The MDSA and MDSANF perform the best at voltage supply larger than 0.6 V and the delay improves as V_{DD} is increased. The MDSA and MDSANF have nearly same read delay at all voltages except 0.5 V and 1 V where MDSA is faster by 22.05% and 20.32% respectively.

2.4.3 POWER CONSUMPTION AND LEAKAGE POWER

The power consumption of MDSA and MDSANF is compared to voltage-mode and current-mode SA architectures and the simulation results are shown in Figure 2.10 (a). As the supply voltage increases, the power consumption of all SAs also increases, which aligns with the expected relationship between power, supply voltage, and current. The power consumption of the MDSA and MDSANF is moderate compared to other SA designs. Among the SA designs, the MHSA consistently exhibits the highest power consumption. This can be attributed to its larger number of transistors, particularly the current conveyor and data line coupling devices, which contribute significantly to power consumption. The inclusion of additional control signals in the MHSA design further adds to the power consumption. In contrast, the CLSA and DLSA designs aim to minimize power consumption by reducing the charge and discharge of the bit line capacitance, resulting in lower power requirements. The presence of pre-charge bit lines also contributes



(a)



(b)

Figure 2.10: (a) Power Consumption and (b) Leakage power comparison of all SAs as a function of the supply voltage (C_{BL} = 50 fF)

significantly to power consumption in SA designs. The DTSA and MCSA have the lowest power consumption with averages of 2.86 μW and 2.03 μW respectively. When comparing to MCSA, MDSA shows a 3.27 μW increase while MDSANF shows an increment of 3.57 μW at V_{DD} = 0.8 V. However, the MDSA and MDSANF consume slightly higher power compared amongst the conventional SAs. Consequently, there is always a trade-off between sensing delay and power consumption when selecting the SA to perform SRAM read operation. Static power, which contributes significantly to the total power consumption of a circuit, is primarily

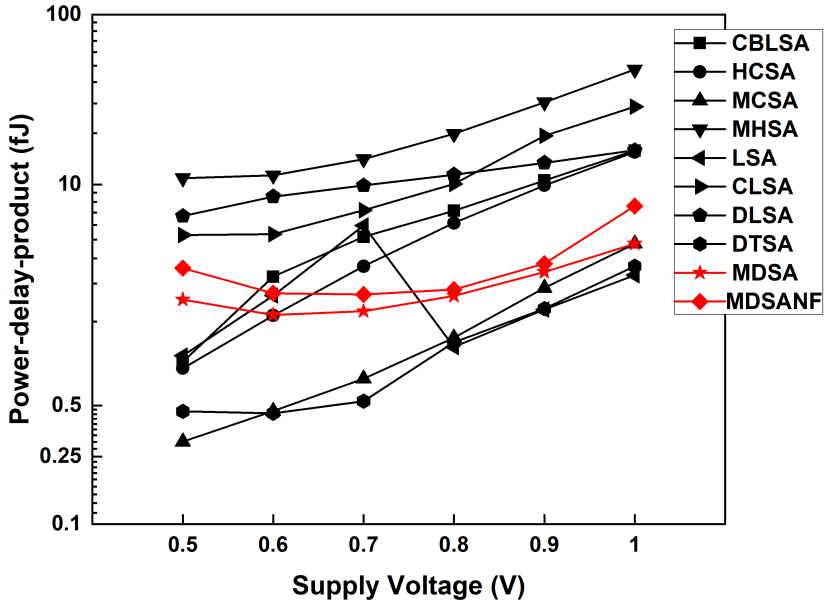


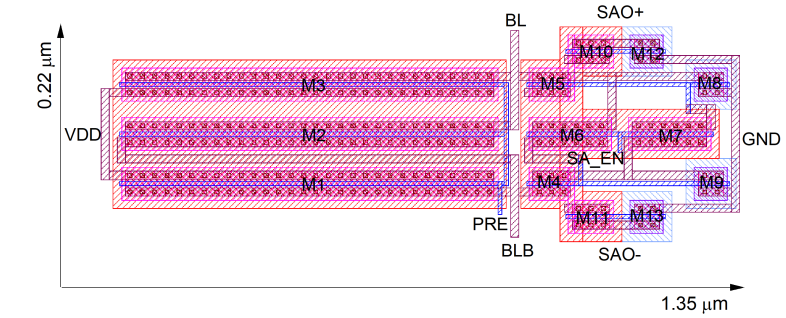
Figure 2.11: Power-delay-product comparison of all SAs as a function of the supply voltage (C_{BL} = 50 fF)

caused by leakage current. It arises due to various factors such as sub-threshold current, gate leakage current, and reverse junction current in a MOS circuit. Mathematically, static power is determined by multiplying the supply voltage (V_{DD}) with the leakage current.
Leakage current refers to the undesired flow of current in a transistor when it is in an idle or off state. The sub-threshold current occurs when there is a small voltage across the transistor, and it results from the imperfect blocking of current by the transistor's gate. The gate leakage current, on the other hand, happens due to the thin insulating layer of the transistor, which allows some current to pass through. Lastly, the reverse junction current occurs as a result of the inherent characteristics of the transistor's junctions. The leakage power for SAs is shown in Figure 2.10 (b). The MDSANF utilizes the power gating effect of the NMOS foot-switch in order to decrease leakage power as compared to MDSA.

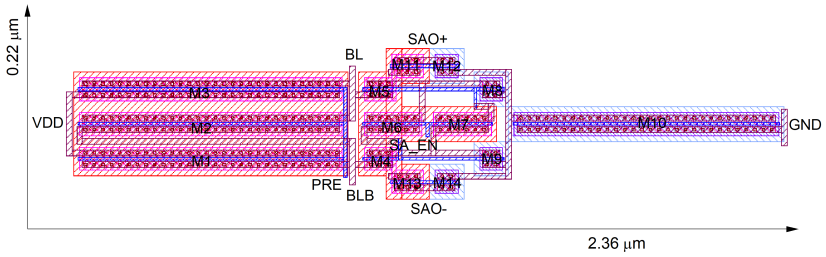
The Power-delay-product (PDP) of all SAs, assuming a bit line capacitance (C_{BL}) of 50 fF, is shown in Figure 2.11. The MHSA consistently exhibits the highest PDP, with an average of 22.35 fJ. The proposed SA design shows a higher PDP than only for DTSA and MCSA, which are both characterized by their simple cross-coupled inverter-based architecture and the low transistor count. Therefore, the MDSA and MDSANF can be employed in faster sensing applications, albeit with a slightly higher power consumption.

2.4.4 AREA CONSIDERATIONS

The selection of a differential SA topology for performing SRAM read is a critical step as a single SA is used to read the data for a column of bit cells. Consequently, the choice of SA topology significantly impacts the footprint of the memory array. Considering that SRAM density is



(a)



(b)

Figure 2.12: Layout design for area estimation of (a) MDSA and (b) MDSANF

Table 2.1: Performance summary of conventional SA designs compared with MDSA and MDSANF at 0.8 V supply voltage

SA Design	Read Delay (ps)	Power Consumption (µW)	PDP (fJ)	Normalized Area
CBLSA	1105.92	6.33	7.00	1.40
HCSA	761.03	7.80	5.93	2.55
MCSA	561.84	2.23	1.25	1.30
MHSA	644.23	30.81	19.85	3.30
LSA	580.81	1.90	1.10	1.00
CLSA	645.54	15.57	10.05	1.16
DLSA	706.55	16.16	11.41	1.65
DTSA	689.15	1.70	1.17	2.22
MDSA	**401.13**	**5.50**	**2.21**	**1.25**
MDSANF	**415.00**	**5.80**	**2.41**	**1.50**

governed by the employed SA, careful consideration of area becomes a critical aspect when selecting the appropriate sensing scheme. The layout of MDSA and MDSANF is specifically designed to estimate the area requirements of the corresponding SA as illustrated in Figure 2.12 (a) and 2.12 (b) respectively. The estimated area for MDSA is 0.297 μm^2 whereas MDSANF occupies 0.5192 μm^2. The area increase for MDSANF is due to the larger sized NMOS foot-switch.The performance summary for MDSA and MDSANF together with conventional SA designs is given in Table 2.1. The estimated area for all SAs is normalized to that of LSA, which is the earliest iterations of SA built around the inverter-latch. The MHSA has largest area due to the large sized PMOS used to build the current conveyor whereas the dual tail sensing scheme used in DTSA is a major contributor to its larger area footprint.

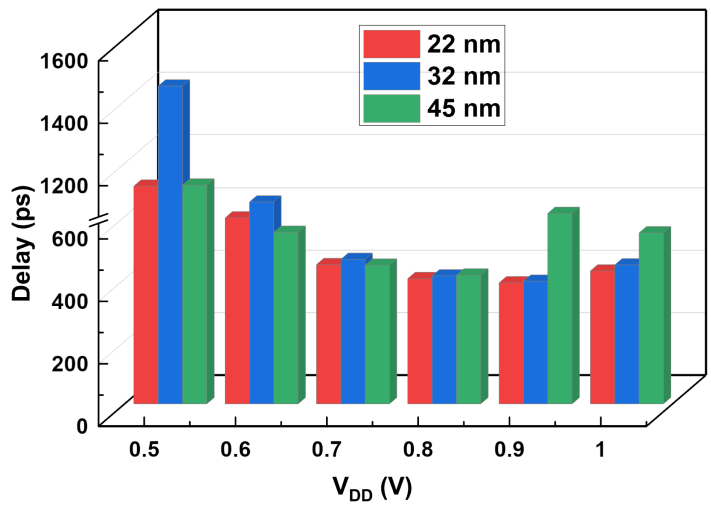
2.4.5 VARIATION–TOLERANCE IMPLICATIONS OF PORTING FROM 22 NM TO 32 NM

Porting the modified decoupled sense amplifier (MDSA) from a 22 nm FinFET process to a 32 nm planar CMOS node entails more than a simple reversal of scaling; it reshapes both the nominal performance envelopes and the underlying variation characteristics. In particular, the threshold-voltage mismatch parameter, $\sigma_{V_T} \approx A_{VT}/\sqrt{WL}$, benefits from the increased channel area WL at 32 nm, even as the planar device's Pelgrom coefficient A_{VT} rises modestly. Consequently, σ_{V_T} falls by approximately 25–30 % when moving from 22 nm FinFETs to 32 nm planar transistors, yielding (i) reduced input-referred offset, (ii) narrower Monte Carlo delay histograms, and (iii) lower tail-fail probabilities in near-threshold operation.

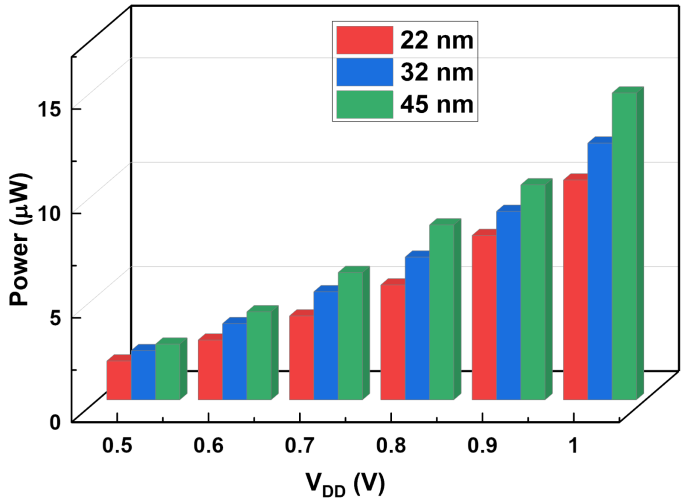
Figure 2.13(a) shows that at an ultra-low $V_{DD} = 0.5$ V, the 32 nm design incurs the highest delay of ≈ 1.52 ns, compared with ≈ 1.22 ns for 22 nm and ≈ 1.24 ns for 45 nm. This penalty stems from reduced drive current and increased gate capacitance in planar devices operating in the subthreshold regime. As V_{DD} rises, the relative difference narrows sharply: at 0.7 V, the 32 nm delay (≈ 0.53 ns) lies only 5% above the 22 nm FinFET delay (≈ 0.50 ns), and by 0.9 V, all three nodes converge within 10 ps of one another. Thus, beyond the near-threshold region, the speed disadvantage of the 32 nm implementation is largely mitigated.

Conversely, Fig. 2.13(b) shows that dynamic power increases monotonically with V_{DD} for all nodes, but with distinct slopes. At 1.0 V, the 32 nm MDSA consumes ≈ 13.3 μW, which is only 12% higher than the 22 nm FinFET value (≈ 11.8 μW) and 21% lower than the 45 nm value (≈ 16.9 μW). The planar 32 nm device therefore occupies an intermediate position, delivering a favorable balance between capacitance-induced power draw and drive-strength benefits.

The combined power–delay product in Fig. 2.14 shows that the 22 nm FinFET implementation provides the lowest PDP over most of the evaluated supply range. At low supply voltage, it minimizes PDP with a value of ≈ 2.75 fJ below 0.6 V. At $V_{DD} = 0.8$ V, the 22 nm node still gives a lower PDP (≈ 3.2 fJ) than the 32 nm and 45 nm implementations (≈ 3.7 fJ and ≈ 4.6 fJ, respectively). Thus, while the 32 nm implementation improves variation tolerance and provides a useful robustness advantage, the 22 nm FinFET node remains more favorable in terms of PDP for the reported operating points.



(a)



(b)

Figure 2.13: Supply-voltage dependence of (a) read delay and (b) dynamic power for the MDSA implemented in 45, 32, and 22 nm CMOS.

Table 2.2: Monte Carlo Read Delay statistics for MDSA (1000 samples, TT corner, V_{DD} = 1 V)

Node	Mean (ps)	σ (ps)	CoV (%)	Min (ps)	Median (ps)	Max (ps)
22 nm	399.70	24.47	6.12	381.35	389.39	506.91
32 nm	417.66	20.73	4.96	387.19	431.92	436.78
45 nm	552.06	9.24	1.67	399.25	552.13	563.00

Note: CoV = $\sigma/\mu \times 100\%$. Median, minimum, and maximum values drawn directly from the Monte Carlo sample.

Beyond nominal performance, statistical variation metrics (see Table 2.2) underscore the robustness benefit of the 32 nm planar SA. Compared with the 22 nm FinFET version—which

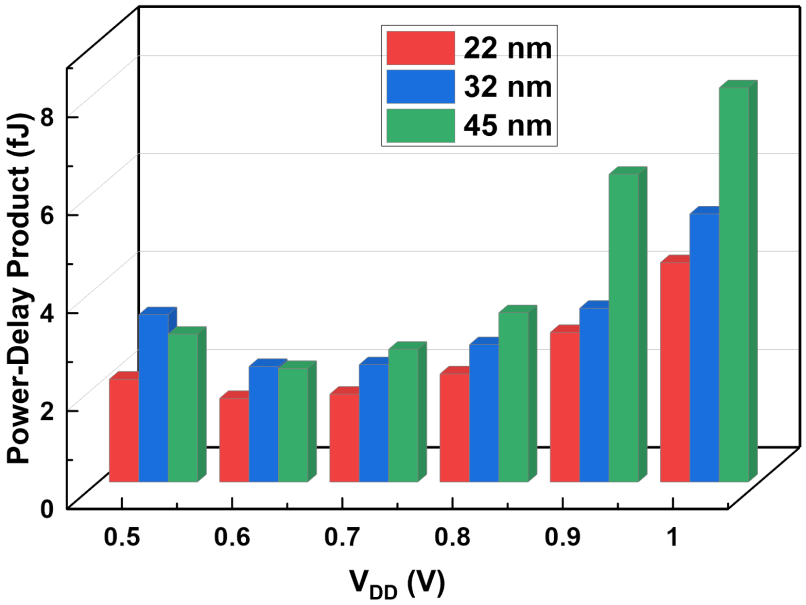


Figure 2.14: Power–delay product (PDP) of the MDSA versus V_{DD} for 45, 32, and 22 nm nodes. The crossover at 0.7 V marks 32 nm as the optimal speed–energy operating point.

exhibits a mean delay of 399.70 ps and a coefficient of variation (CoV) of 6.12%—the 32 nm implementation has a slightly increased mean delay of 417.66 ps but a reduced CoV of 4.96%. Its median delay (431.92 ps) closely matches the mean, and its overall delay range (387.19 ps–436.78 ps) is about 25% narrower than the 22 nm span (381.35 ps–506.91 ps). The 3σ delay spread shrinks from approximately 73.4 ps at 22 nm to 62.2 ps at 32 nm—a 15% reduction—directly improving six-sigma yield under ±30 mV threshold variation. Crucially, this variation tolerance is "free", requiring no additional pre-charge margin or assist circuitry and thus preserving energy efficiency and area.

From a layout perspective, the larger diffusion widths at 32 nm remain confined to the sense-amplifier column when the PMOS/NMOS finger count is kept constant, incurring only a modest area penalty relative to the 22 nm FinFET version. Therefore, the 32 nm port yields an SA that is marginally slower in the near-threshold regime but substantially more variation-tolerant and energy-efficient in the nominal window, achieving an optimal compromise for energy-constrained SRAM macros targeted at IoT and low-power cache applications.

2.5 SUMMARY OF IMPORTANT RESULTS

This chapter presented the design and evaluation of two modified decoupled sense-amplifier structures for low-voltage differential SRAM read operation. The proposed MDSA and MDSANF were developed to improve sensing speed while maintaining variation tolerance, practical power consumption, and acceptable layout overhead. Their behavior was examined under supply-voltage scaling, bit-line loading, process corners, Monte Carlo mismatch, sizing variation, and node portability. The key findings are summarized below:

1. The modified decoupled sense amplifier (MDSA) and the MDSA with NMOS foot-switch (MDSANF) were proposed for differential SRAM sensing at the 22 nm technology node. Both designs decouple the latch from the bit-lines during pre-charge to reduce bit-line disturbance before evaluation.
2. The NMOS foot-switch in MDSANF reduces static conduction paths and provides better control over the evaluation window. This improves leakage and timing control, although it introduces a modest delay and area penalty compared with MDSA.
3. At $V_{\mathrm{DD}} = 0.8$ V and $C_{\mathrm{BL}} = 50$ fF, MDSA achieves read-delay reductions of 28.6 %, 41.79 %, 37.74 %, and 30.94 % over MCSA, DTSA, MHSA, and LSA, respectively. MDSANF gives comparable reductions of 26.13 %, 39.78 %, 35.58 %, and 28.55 %.
4. Delay characterization over the 0.5–1.0 V supply range confirms that both proposed amplifiers maintain faster sensing from near-threshold to nominal operation. This is especially important for low-voltage SRAM read paths where the available bit-line differential is small.
5. Parameter sweeps involving ΔV_{BL}, C_{BL}, and process corners show stable and monotonic delay trends under realistic array loading. These results confirm that the proposed sensing structure remains well behaved across key SRAM operating conditions.
6. Monte Carlo analysis with 10 000 samples and ±30 mV threshold-voltage variation confirms reduced delay spread and improved variation tolerance. The proposed decoupled sensing mechanism therefore improves both nominal speed and mismatch robustness.
7. Sizing analysis identifies an 88 nm width for the decoupling PMOS pair as the preferred design point. This width provides a favorable power–delay trade-off while remaining compatible with SRAM column-pitch constraints.
8. Power, leakage, PDP, and layout results show that the speed improvement is obtained with practical overhead. The MDSA occupies 0.297 $\mu\mathrm{m}^2$, while MDSANF occupies 0.5192 $\mu\mathrm{m}^2$, both acceptable for column-level SRAM sensing.
9. Porting MDSA from 22 nm FinFET to 32 nm planar CMOS incurs only a 4.5 % increase in mean read delay at $V_{DD} = 1$ V. At the same time, the CoV reduces from 6.12 % to 4.96 %, and the delay range narrows by 25 %, indicating improved variation tolerance without additional assist circuitry.

Overall, the results establish decoupled latch sensing as an effective approach for improving SRAM read speed and variation tolerance. MDSA provides the stronger speed advantage, whereas MDSANF adds leakage and evaluation-window control through the NMOS foot-switch. The combined delay, variation, PDP, and layout results show that the proposed amplifiers are practical sensing solutions for low-voltage differential SRAM macros. These designs therefore provide the sensing foundation for the SRAM cell and SRAM-IMC developments presented in the subsequent chapters.

CHAPTER – 3

PROPOSED 8T SRAM CELL DESIGN FOR IMPROVED PERFORMANCE

3.1 OVERVIEW OF THE CHAPTER

The bit cell serves as the cornerstone of SRAM memory design, exerting a direct influence on critical aspects such as speed, power efficiency, area utilization, and reliability. While the conventional 6T SRAM cell has long been the standard due to its compact layout and ease of integration, aggressive technology scaling and the trend toward lower supply voltages have increasingly exposed its limitations—most notably reduced read stability, higher susceptibility to noise, and significant vulnerability to process, voltage, and temperature (PVT) variations. To overcome these issues, researchers have explored alternative architectures, among which the 8T SRAM cell has drawn considerable attention. By decoupling the read and write operations through separate access paths, the 8T design enhances read stability and improves noise immunity. However, single-ended variants, though simpler and more energy efficient, often struggle with robustness, while differential designs offer stability at the expense of added complexity and power overhead. Accordingly, the following objective is defined:

"Performance Enhancement of a modified differential bit cell for energy efficient SRAM"

Methodology

1. Design a variation tolerant single-ended 8T SRAM using multi-threshold transistors for improved performance.
2. Characterize the impact of process variations on the proposed 8T cell for variety of operating conditions.
3. Validate the performance of 8T cell against the pre-existing cells of varied cell structures.
4. Propose a differential 8T cell structure to overcome the shortcomings of single-ended architectures.
5. Compare the proposed single-ended and differential 8T SRAM cells to highlight their architectural differences, performance impact, and sensitivity to process variations.

The structure of this chapter is outlined below. Section 3.3 introduces the proposed single-ended process-induced variation resistant 8TS SRAM cell and evaluates its behavior under voltage scaling in terms of stability, read/write performance, and robustness against process variations. Section 3.4 provides a comparative analysis of the proposed 8TS cell with other existing SRAM architectures, including designs with different transistor counts. Section 3.5 presents the proposed

differential 8T SRAM cell architecture aimed at overcoming the shortcomings of single-ended designs. Section 3.6 discusses the stability margin comparison between the single-ended and differential 8T cells, while Section 3.7 focuses on delay and power analysis for read and write operations. Section 3.8 details the results of Monte Carlo simulations to assess threshold voltage variation tolerance. The concluding section, Section 3.9, presents a summary of the important results.

3.2 INTRODUCTION

The relentless march of technology is driving a surge in smart devices and the widespread adoption of the Internet of Things (IoT) across diverse industries. This trend necessitates the development of increasingly energy-efficient architectures to power these advancements. One key application domain is wireless sensor networks (WSNs), which continuously monitor various environmental parameters using sensors and wireless transmission of data. However, this continuous operation comes at a significant energy cost, demanding extended battery life for these sensors [147–150]. The need for faster and less powerful devices has pushed digital designers into exploring deeper in the realm of SRAM [151, 152]. Similarly, space applications like satellite communication, crucial for weather forecasting, military surveillance, and other critical functionalities, require low-cost, lightweight satellites with extended battery life and the ability to withstand harsh space radiation [153]. Some focus has also been put on the memory peripherals like the sense amplifier and their optimization in order to improve the memory array performance. In these applications, high-density memories play a vital role, but they face the challenge of soft errors caused by high-energy particles like alpha (α) and beta (β) particles that can get trapped within semiconductor devices, corrupting stored data and jeopardizing circuit performance. Addressing these challenges, recent research in SRAM bitcell design has explored the potential of bit-interleaving architectures as a promising solution to mitigate soft errors. This technique offers a compelling approach to enhance the reliability of memory devices in these demanding applications.

Over the years, the literature has introduced various configurations of SRAM cells with the aim of achieving superior performance compared to conventional designs[154]. Standard 8T SRAM cell improves Read Stability Noise Margin by employing isolated read path through the inclusion of two additional transistors. However, this benefit comes with a trade-off of increased read path leakage, which worsens with technology miniaturization. To address this, [155] proposes a modified isolated read path design within their SRAM cell. This modification specifically tackles data-independent leakage, thereby mitigating overall leakage and consequently improving RSNM. [156] introduced a novel SRAM cell design that integrates a Schmitt-trigger inverter to enhance both read and write stability margins. This innovative approach capitalizes on the robust noise margin properties inherent in the voltage transfer curve of the Schmitt-trigger, thereby improving noise immunity. However, it's worth noting that the inclusion of the Schmitt-trigger INV leads to a larger cell footprint. Multiple studies [157–160] have explored the adoption of single bitline architectures in SRAM cells as a strategy to lower power consumption. This

approach centers around reducing the bitline switching activity factor (α) below 50% during read/write operations. However, single-ended designs lacking write-assist techniques suffers from significant degradation in their ability to write '1' data, often resulting in write failures. To overcome this issue, several researchers [161, 162] proposed incorporating a feedback-cutting NMOS transistor or transmission gate within the cell core. This mechanism disrupts the feedback path of the cross-coupled inverters during write operations, leading to an improvement in WM. An alternative approach to mitigate write '1' difficulties in single-ended SRAM cells involves the data-aware power cutoff write-assist mechanism. The 11T cell proposed in [163] employs this technique to achieve improved WM. However, it comes at the expense of increased active power consumption attributed to the introduction of additional capacitance on the bitline. [164] proposed a bit-interleaving 12T SRAM cell which utilizes a read buffer to enhance RSNM and employs a combination of cross-point data-aware write word lines and a power-cutoff write-assist technique. This combination improves WM and eliminates write half-select issues. However, the dual-bitline structure inherent to the design leads to higher dynamic power consumption.

3.3 PROPOSED SINGLE-ENDED PROCESS-INDUCED VARIATION RESISTANT 8TS SRAM CELL

The proposed single-ended, single-port 8TS SRAM architecture introduces a novel design that enhances performance, improves reliability, and minimizes leakage power consumption. The bit-cell comprises eight MOS transistors, including a feedback control transistor (M_5) integrated between where conventionally two inverters reside. However, here we have the gate of (M_4) pulled to GND while the (M_3) and (M_6) transistors form the singular inverter. This structure allows formation of the storage nodes Q and QB, which hold the data. The design also incorporates dedicated read and write ports for efficient data transfer, as shown in Figure 3.1(A). The write bitline (WBL) is connected to the storage node Q via an access transistor (M_1), while the read bitline (RBL) utilizes stacked NMOS transistors (M_7 and M_8) to transfer data from the storage node QB. This separation of read and write paths ensures stable operation under varying conditions.

During a write operation, the feedback control transistor (M_5) is turned off by deactivating the write signal (WL), disconnecting the cross-coupled inverters. This allows the new data to overwrite the stored value without interference. The write process begins by precharging WBL to V_{DD}. While performing write '1', WBL remains at V_{DD}, driving Q to V_{DD} and QB to GND through M_6. Conversely, when writing a '0', WBL is discharged to GND, pulling Q low and QB high. Once the data is written, the feedback control transistor (M_5) is turned back on, the inverters are reconnected which stabilizes the stored data. Figure 3.1(C) shows the transient simulation during a write operation, showing the dynamic behavior of input control signals and their impact on the storage nodes.

In the hold mode, the feedback connection ensures data stability. The pull-down NMOS transistors (M_4 and M_6) in the structure are optimized to remain in a cut-off state, reducing leakage and preventing unintended switching. High-V_T PMOS transistors (M_1 and M_3) are

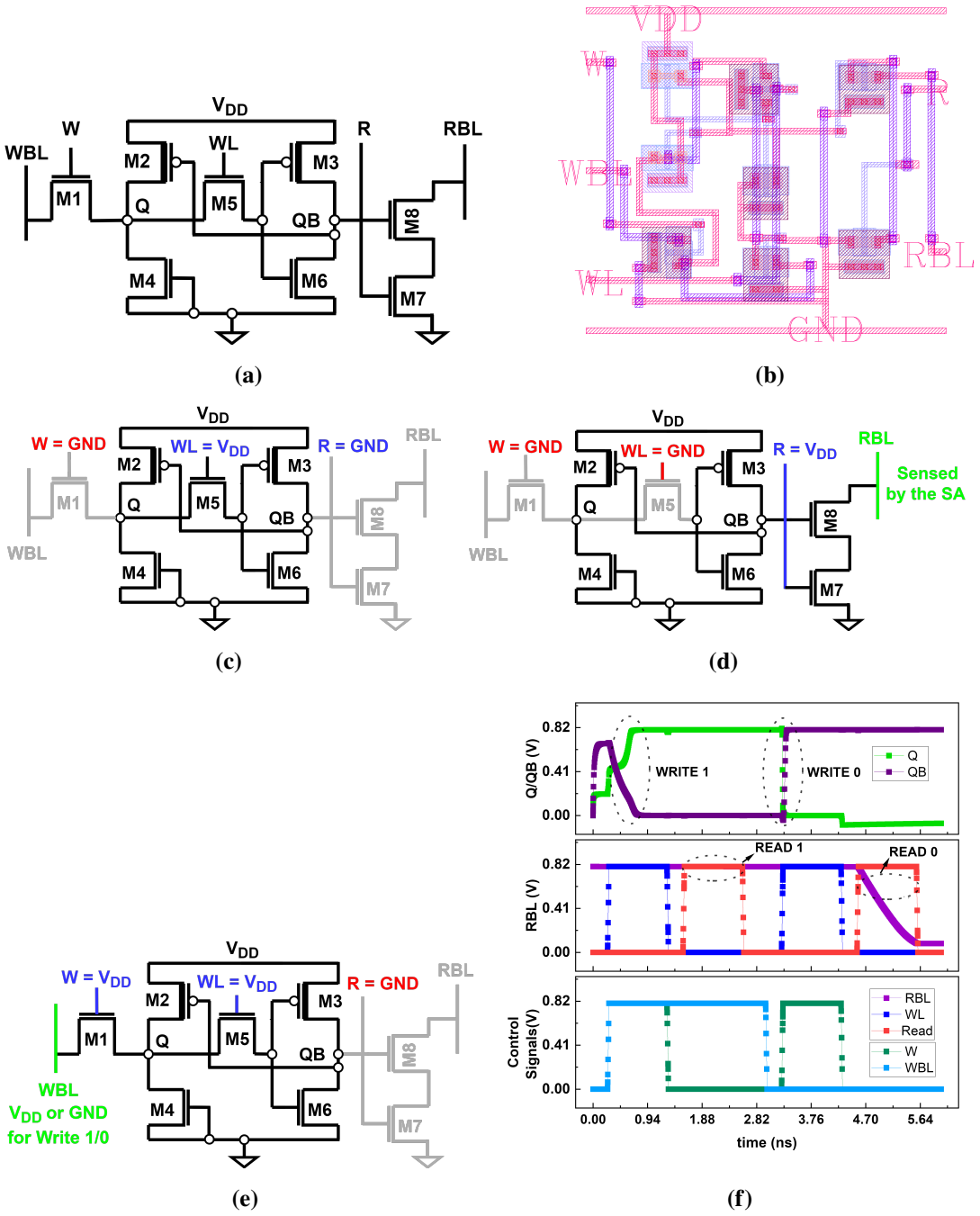


Figure 3.1: Proposed single-ended, single port 8T SRAM bitcell (A) 8TS circuit schematic, (B) layout of the 8TS bit-cell; Equivalent circuit for (C) hold mode, (D) read mode, and (E) write mode; (F) transient simulation with control signals for all the operations of 8TS Cell.

employed to minimize leakage current. By employing high-V_T transistors, the leakage power, expressed as $P_{\text{leakage}} = I_{\text{OFF}} \cdot V_{\text{DD}}$, is significantly reduced. The leakage current decreases exponentially as V_T increases which in turn leads to reduced leakage. This design ensures low static power consumption, making it suitable for energy-efficient applications.

The read operation utilizes a dedicated read port, ensuring that the stored data remains undisturbed. The read bitline (RBL) is precharged to V_{DD} using peripheral circuitry. When the

read signal (R) is activated, the state of QB determines the conductance of M_8. If QB is '1', M_8 turns on, pulling RBL to ground through M_7 for R at V_{DD}. If QB is '0', M_8 remains off, leaving RBL at V_{DD}. The voltage on RBL is then sensed by a sense amplifier to determine the stored data. Figure 3.1(D) illustrates the transient simulation of the read operation, showing the stable behavior of the storage nodes and the read bitline.

The proposed architecture offers several advantages over conventional designs. By selectively disconnecting the feedback loop during write operations, the design improves the write margin, ensuring reliable data storage even under process variations. The dedicated read port enhances stability by decoupling the read and write paths, reducing the risk of read disturbance. The use of high-V_T PMOS transistors significantly reduces leakage current. Additionally, Monte Carlo simulations demonstrate the design's robustness under process, voltage, and temperature (PVT) variations, making it highly suitable for low-power and high-reliability applications.

3.3.1 STABILITY ANALYSIS OF THE 8TS SRAM CELL

The stability of an SRAM bit-cell is crucial for maintaining data integrity and preventing bit flips caused by noise or environmental factors. Static noise margin (SNM) is a widely used metric to evaluate stability, representing the maximum DC noise voltage that the cell can tolerate without flipping its stored data. The SNM is determined by the side length of the largest square that can fit within the smaller lobe of the butterfly curve, which is derived by overlapping the voltage transfer characteristics (VTC) of the cross-coupled inverters ($INV1$ and $INV2$) in the cell. The butterfly curve for the proposed 8TS SRAM bit-cell is shown in Figure 3.2(A), where the SNM value is 0.23242 V at $V_{DD} = 0.8$ V. The SNM is same for hold and read SNM. This equivalence is due to the activation of similar transistors during both operations, making the cell effectively read SNM-free [165].

To ensure robustness under process variations and noise, a Monte Carlo analysis with 2000 samples was performed at $V_{DD} = 0.8$ V, as illustrated in Figure 3.2(B). This analysis accounted for a 10% tolerance in the threshold voltage (V_{TH}) of the transistors. The results show that the best-case and worst-case SNM values are 0.26905 V and 0.19583 V, respectively. The distribution of SNM values highlights the cell's reliability under varying conditions, demonstrating a 6-σ deviation, which ensures robust performance in large-scale memory arrays.

The dependence of SNM on the supply voltage (V_{DD}) was also investigated, as shown in Figure 3.2(C). The SNM was calculated for V_{DD} values ranging from 0.6 V to 1.2 V. The results reveal a positive correlation between V_{DD} and SNM, with higher supply voltages leading to increased stability. Specifically, the SNM values observed at V_{DD} values in volts of 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, and 1.2 are 0.15, 0.19, 0.23, 0.27, 0.30, 0.32, and 0.34, respectively. This trend is consistent with the expected behavior, as a higher V_{DD} improves the noise margins of the cross-coupled inverters by increasing the voltage difference between the logic levels.

Temperature dependence is another critical factor affecting the SNM of the SRAM bit-cell. Figure 3.2(d) illustrates the variation of SNM with temperature for different V_{DD} values. The simulated results show that the proposed 8TS cell maintains a stable noise margin over the

evaluated temperature range, with only a modest temperature dependence. For example, at V_{DD} = 0.8 V, the SNM increases slightly from approximately 0.23 V at −25°C to approximately 0.245 V at 100°C. This indicates that, for the proposed sizing and operating conditions, the temperature-induced shift in the device characteristics does not degrade the butterfly-curve stability margin.

The observed trend is governed by the combined effect of threshold-voltage shift, mobility variation, leakage current, and the resulting movement of the inverter switching points with temperature. Since these effects influence the pull-up, pull-down, and access/read devices differently, the SNM variation with temperature need not be strictly monotonic for all SRAM

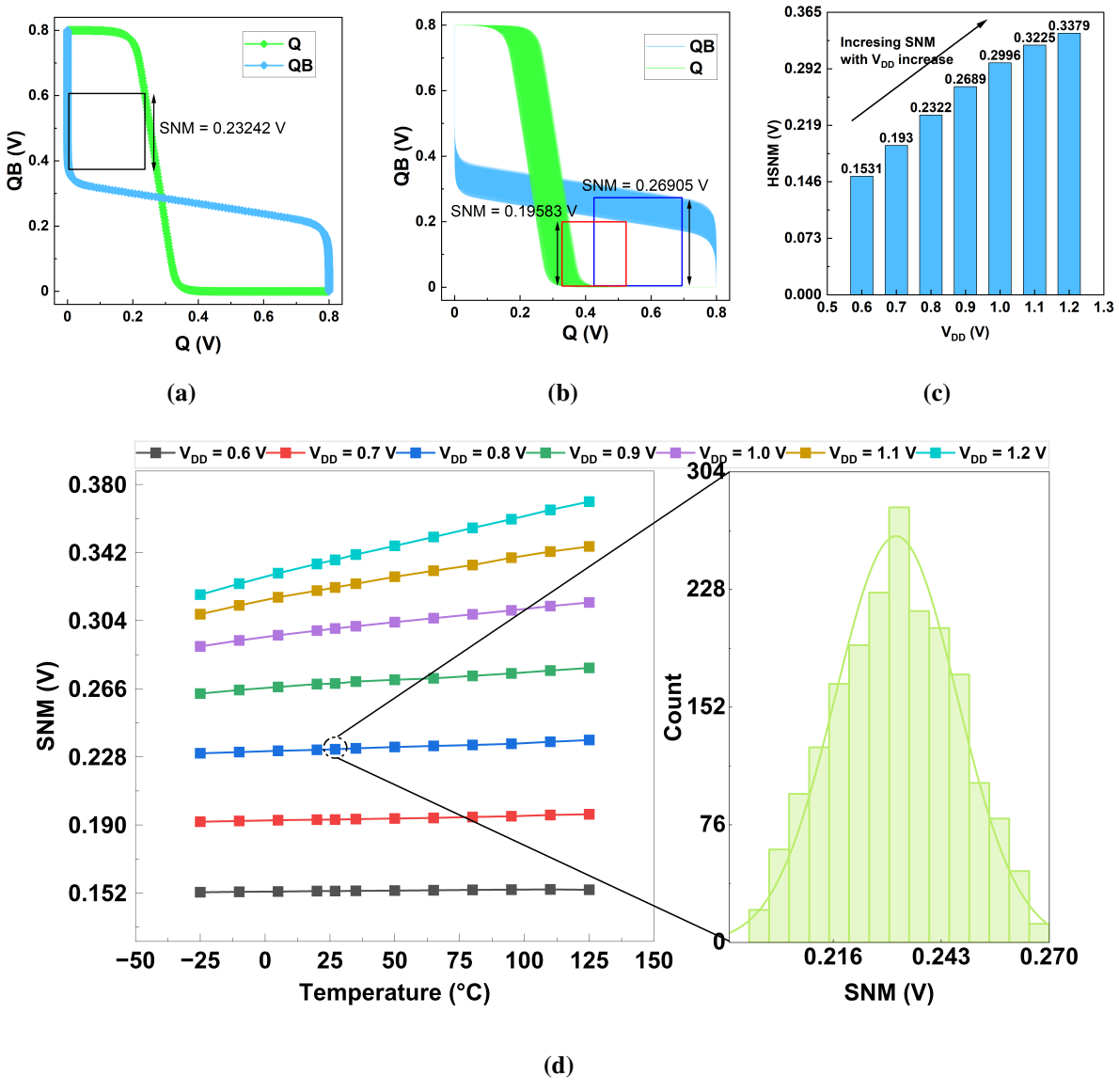


Figure 3.2: Static noise margin butterfly curve of the 8TS SRAM cell for (A) SNM calculation, (B) butterfly curve for Monte Carlo (2000-point) analysis at V_{DD} = 0.8 V, (C) SNM as a function of supply voltage, and (D) SNM as a function of temperature (Inset: distribution of SNM at T = 27°C for V_{DD} = 0.8 V)

topologies and supply voltages. In the proposed 8TS cell, the extracted SNM remains adequate across the evaluated temperature range, confirming stable operation under temperature variation.

The inset in Figure 3.2(d) presents the SNM distribution at $V_{DD} = 0.8$ V and $T = 27°C$, based on a 2000-point Monte Carlo simulation. The distribution exhibits a mean SNM of approximately 0.231 V with a standard deviation of approximately 0.015 V, further demonstrating the robustness of the proposed 8TS SRAM bit-cell under realistic process-induced variation.

The stability analysis of the proposed 8TS SRAM bit-cell highlights its resilience under supply-voltage, temperature, and process variations. The positive correlation between V_{DD} and SNM confirms improved stability at higher supply voltages, while the Monte Carlo analysis verifies robustness under threshold-voltage variation. The temperature sweep further shows that the proposed cell maintains adequate SNM across the evaluated temperature range, making it suitable for energy-efficient SRAM operation under practical operating conditions.

3.3.2 WRITE MARGIN ANALYSIS OF 8TS SRAM CELL

The write margin (WM), or write ability, of an SRAM bit-cell is a critical metric that measures the cell's stability during a write operation. It quantifies the cell's ability to reliably flip its stored data when required while resisting undesired flipping due to noise or voltage fluctuations. The WM is defined as the difference between V_{DD} and the write trip point (WTP), where the WTP is the voltage at which the storage nodes Q and QB flip states [166]. As illustrated in Figure 3.3 (a) and Figure 3.3 (b), the WTP determines the voltage threshold at which the data stored in the cell changes. For the proposed 8TS SRAM bit-cell, the WM was calculated to be 0.5386 V and 0.5153 V for write-0 and write-1 operations respectively, which is higher than the ideal value of $0.5 \cdot V_{DD}$ (0.4 V at $V_{DD} = 0.8$ V). This demonstrates that the proposed cell achieves moderately good write ability, striking a balance between preventing accidental flipping and enabling reliable data writing.

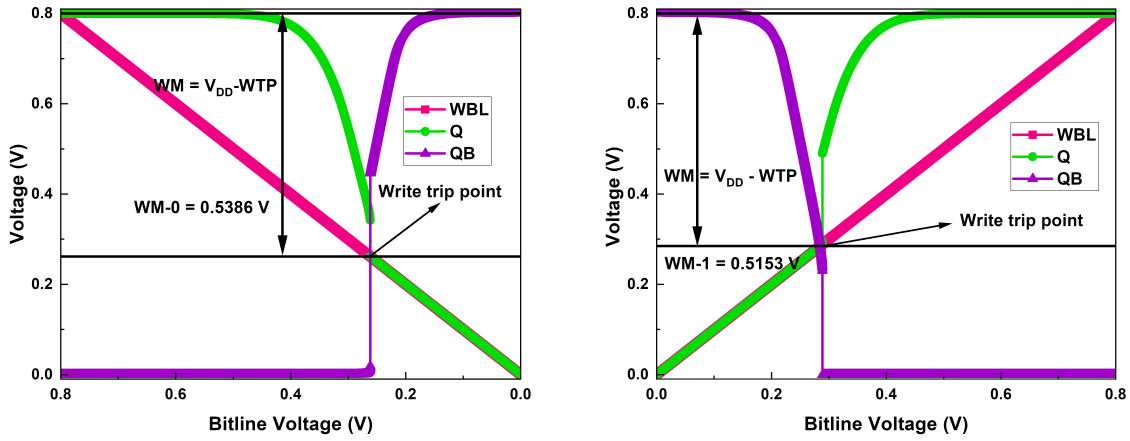
A low WM would make the cell susceptible to flipping at low bit-line (BL) voltages or due to voltage variations and noise, whereas a high WM would make it difficult to write data into the cell. The decoupled feedback loop in the proposed 8TS SRAM bit-cell enhances the WM by reducing the hold strength of the inverter during write operations. This design allows the pass transistors to flip the stored data more efficiently, thereby improving write ability. Notably, the mismatched write margins for two states are a result of the deliberate asymmetry incorporated into the cell design, as shown in Figure 3.3 (a) and Figure 3.3 (b). This ensures robust and predictable performance under varying process, voltage, and temperature (PVT) conditions.

Figure 3.3 (c) illustrates the impact of supply voltage scaling on the WM. As V_{DD} increases from 0.6 V to 1.0 V, the WM also increases, highlighting the positive correlation between supply voltage and write ability. The proposed cell's ability to maintain a stable WM across a wide range of supply voltages further demonstrates its robustness and suitability for energy-efficient SRAM applications.

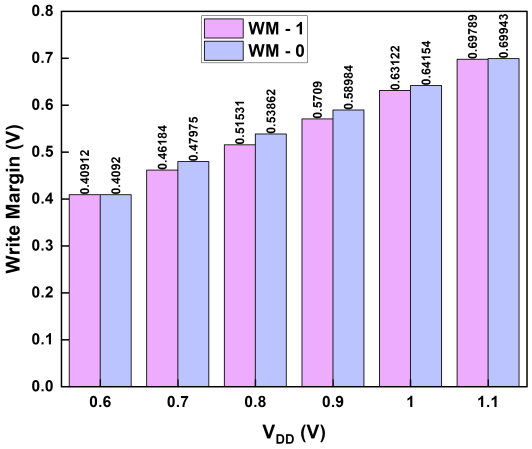
During the fabrication of integrated circuits, process variations often arise due to manufacturing imperfections, leading to deviations in transistor parameters such as threshold voltage (V_T).

To assess the robustness of the proposed 8TS SRAM bit-cell under such variations, a Monte Carlo analysis was conducted, as shown in Figure 3.4. The analysis considers variations across three standard process corners: SS (slow NMOS, slow PMOS), TT (typical),SF (slow NMOS, ast PMOS), FS (fast NMOS, slow PMOS) and FF (fast NMOS, fast PMOS), assuming a V_T variation of 30 mV. Additionally, the analysis accounts for supply voltage variations, with ΔV_{DD} set to 25% of the nominal V_{DD} (0.8 V).

Figure 3.4 (a) through 3.4 (o) show the WM distributions at $V_{\mathrm{DD}} - \Delta V_{\mathrm{DD}}$, V_{DD} and $V_{\mathrm{DD}} + \Delta V_{\mathrm{DD}}$, respectively, for the SS, TT, FF, FS and SF process corners. The mean WM values for each process corner are also highlighted. For the TT corner, the mean WM value (0.51654 V) closely aligns with the theoretical WM value (0.51531 V) at $V_{\mathrm{DD}} = 0.8\,\mathrm{V}$. In the SS corner, the mean WM value is slightly lower (0.51001 V), while in the FF corner, the WM value is slightly higher (0.5257 V) due to the increased V_{T}. These results confirm the cell's ability to maintain stable performance across varying process conditions.



(a) **(b)**



(c)

Figure 3.3: Write Margin's calculation for the proposed 8TS SRAM bit-cell (A) WM-0, (B) WM-1, and (C) the impact of supply voltage scaling.

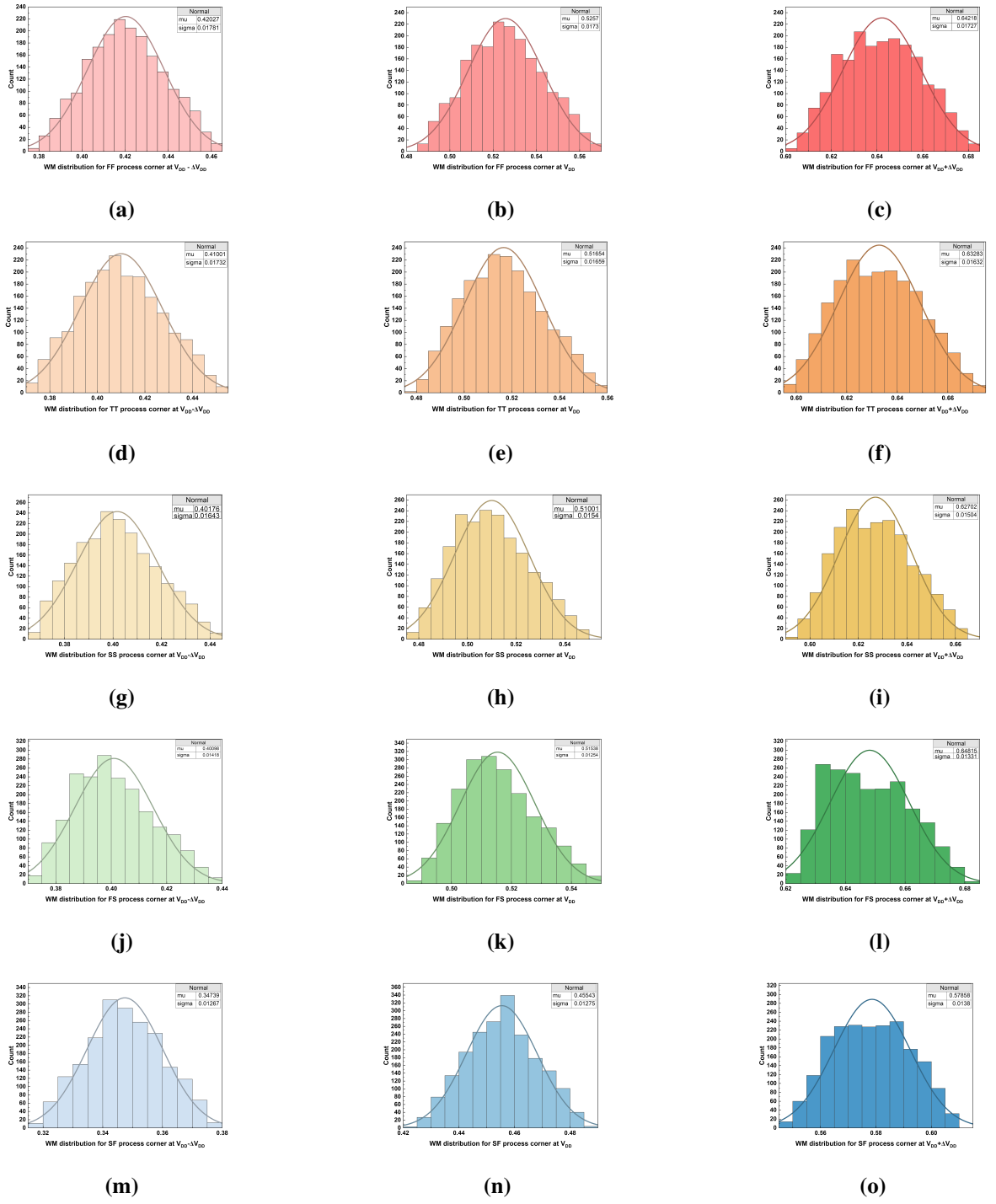


Figure 3.4: Monte Carlo simulations of the proposed 8TS cell across process corners, evaluating write margin (WM) under threshold voltage V_{TH} variations of ±30 mV and supply voltage fluctuations ΔV_{DD} of 25% around the nominal 0.8 V.

The trend observed in 3.4 (a–o) indicates that the write-margin (WM) decreases as we move from the FF corner through TT and SS to the skewed FS and SF corners for any fixed supply voltage. This behavior arises because the FF corner's lower threshold voltage weakens the inverter hold strength, easing writes, whereas SS's higher threshold strengthens hold and raises

Table 3.1: Coefficient of Variance (CoV) Analysis for Write Margin (WM) across Process Corners and Supply Voltages.

Process Corner	Supply Voltage Condition	Mean WM (V)	CoV (%)
SS	$V_{DD} - \Delta V_{DD}$	0.40176	4.09
	V_{DD}	0.51001	3.02
	$V_{DD} + \Delta V_{DD}$	0.62702	2.39
TT	$V_{DD} - \Delta V_{DD}$	0.41001	4.22
	V_{DD}	0.51654	3.21
	$V_{DD} + \Delta V_{DD}$	0.63283	2.58
FF	$V_{DD} - \Delta V_{DD}$	0.42027	4.24
	V_{DD}	0.52570	3.29
	$V_{DD} + \Delta V_{DD}$	0.64218	2.69
FS	$V_{DD} - \Delta V_{DD}$	0.40098	3.54
	V_{DD}	0.51538	2.43
	$V_{DD} + \Delta V_{DD}$	0.64815	2.05
SF	$V_{DD} - \Delta V_{DD}$	0.34739	3.65
	V_{DD}	0.45543	2.80
	$V_{DD} + \Delta V_{DD}$	0.57858	2.38

WM; the FS and SF curves fall neatly between these extremes. Conversely, WM increases with higher V_{DD}, as shown in Figure 3.3(c).

The Monte Carlo simulations in Figure 3.4(A) through Figure 3.4(O) further illustrate the WM distributions under process variations at $V_{DD} = 0.8\,V$. The WM distribution for each process corner demonstrates a small standard deviation, confirming the proposed cell's robustness to threshold voltage and supply voltage variations. The FF corner exhibits the highest deviation due to the lower V_T, while the SS corner shows the lowest deviation, consistent with the higher V_T in this corner.

The write margin analysis of the proposed 8TS SRAM bit-cell demonstrates its superior write ability and robustness under varying supply voltages and process conditions. The decoupled feedback loop design significantly enhances the WM, ensuring reliable write operations even under adverse conditions. The Monte Carlo analysis confirms the cell's stability and predictable performance, making it a strong candidate for energy-efficient and reliable SRAM applications.

The coefficient of variance (CoV) is a statistical measure that represents the ratio of the standard deviation to the mean, expressed as a percentage. It provides a normalized measure of the dispersion of a dataset, allowing for a comparative analysis of variability across different process conditions and supply voltages. For the Monte Carlo simulations of the proposed 8TS

SRAM bit-cell, the CoV was calculated for the write margin (WM) distributions at five process corners—SS, TT, FS, SF and FF—under three different supply voltages: $V_{DD} - \Delta V_{DD}$, V_{DD}, and $V_{DD} + \Delta V_{DD}$.

From Figure 3.4, the CoV shows distinct trends across process corners and voltage variations. At $V_{DD} = 0.8\,V$, the TT corner exhibits the lowest CoV (indicating minimal variability), suggesting stable and predictable performance. Conversely, the FF corner shows the highest CoV, reflecting greater variability due to reduced threshold voltage (V_T) in fast transistors. As supply voltage increases from $V_{DD} - \Delta V_{DD}$ to $V_{DD} + \Delta V_{DD}$, the CoV decreases across all process corners, indicating improved stability at higher voltages. The CoV for WM values across process corners is compiled in Table 3.1. This behavior aligns with the theoretical expectation that higher supply voltages provide greater write margin, reducing sensitivity to process variations.

3.3.3 POWER CONSUMPTION AND LEAKAGE ANALYSIS OF 8TS SRAM CELL

A significant portion of an SoC's power consumption is attributed to SRAM bit-cells. Reducing the power dissipated by these cells can significantly lower the overall energy consumption of the system. The power dissipation of an SRAM bit-cell is categorized into three main types, depending on its operational state: write power, read power, and leakage power. Write power corresponds to the energy used during data writing, while read power accounts for the energy required to retrieve stored data. Leakage power, on the other hand, is the power dissipated when the cell is in a standby or hold state.

The variation in read and write power for the 8TS SRAM cell across different supply voltages is illustrated in Figure 3.5(A). The figure presents the power dissipation trends for read and write operations as V_{DD} increases from 0.6 V to 1.2 V. It can be observed that both write-1, write-0 power as well as read power consumption increase with rising V_{DD}, this can be clearly seen as power consumption is directly related to the supply voltage. At V_{DD} = 0.8 V, the power dissipation during the read operation is 0.186 μW, while for the write-1 and write-0 operation, it is 0.247 μW and 8.89 μW. During a read cycle, additional power is also consumed by the sense amplifier and peripheral circuitry. However, these are not accounted for in our current discussion.

Figure 3.5(B) highlights the influence of bitline capacitance (C_{BL}) on leakage power across different supply voltages. The metal lines for a conventional 6T array result in 50fF bitline capacitance [167]. We have varied the C_{BL} from a few to a few hundred fF in order to showcase the impact of changing the array size on leakage power, which is a crucial factor in SRAM power dissipation, especially in low-power applications. At V_{DD} = 0.8 V, leakage power is observed to be around 2.52 nW (C_{BL} = 50fF), with a noticeable trend of increasing leakage power as C_{BL} decreases. The leakage power decreases by only 59.62% for a 100× increase in C_{BL}. This trend is attributed to the greater capacitive load on the bitlines, which affects the leakage characteristics of the transistors within the SRAM cell. A larger C_{BL} results in reduced leakage currents due to higher electric fields, particularly influencing subthreshold and gate leakage components. However, beyond a certain capacitance value, the rate of leakage power increase tends to saturate

as the dominant leakage component transitions from subthreshold leakage to gate oxide leakage. Since leakage power is proportional to $V_{DD}{}^2$, even small increases in V_{DD} lead to a significant rise in leakage dissipation.

The write-1 operation consumes 1.08 μW, whereas the write-0 operation dissipates 26.61 μW. The significantly higher write-0 power is due to the discharge of large capacitive bitlines to ground, while write-1 only requires charging the bitlines to V_{DD}. During a write-0 event, there is a period where both transistors in the inverter core are partially conducting, leading to a rise in short-circuit current, which contributes to increased power dissipation. In contrast, during a write-1 operation, the PMOS transistor turns on quickly while the NMOS turns off rapidly, reducing the overlap period and minimizing short-circuit power loss. Additionally, the power dissipation of an SRAM bit-cell exhibits an inverse relationship with temperature. As temperature increases, power consumption decreases, as depicted in Figure 3.5(C). The figure

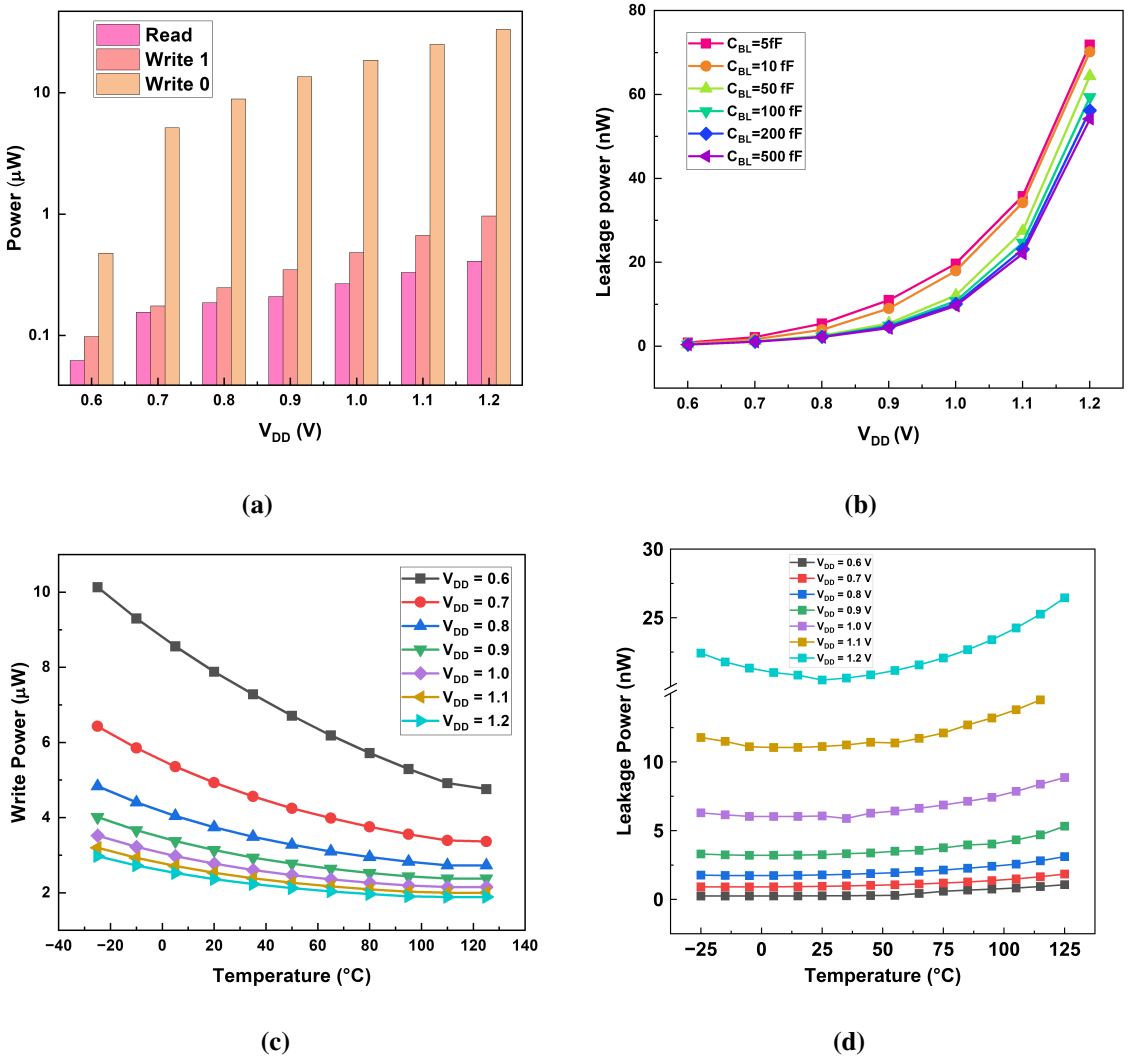


Figure 3.5: Power consumption of the 8TS SRAM cell as a function of supply voltage: (A) read and write power variation, (B) leakage power as a function of bitline capacitance, and (C) temperature dependence of write power and (D) leakage power at multiple V_{DD}.

illustrates how power varies with temperature while V_{DD} is incremented from 0.6 V to 1.2 V in steps of 0.1 V. This decline in power dissipation with increasing temperature is primarily due to the reduction in carrier mobility, which lowers the drive current of the transistors, thereby decreasing dynamic and leakage power components. However, at very high temperatures, the threshold voltage degradation and increased subthreshold leakage can introduce additional leakage components, leading to a possible rise in overall power dissipation.

3.3.4 READ AND WRITE DELAY ANALYSIS OF 8TS SRAM CELL

The delay associated with read and write operations in SRAM cells is a key performance metric that affects overall memory access time. To evaluate the robustness of the proposed 8TS SRAM cell, we analyze its read and write delays under different temperature conditions. Read delay is defined as the time elapsed from the activation of the read signal until the sense amplifier output reaches half of V_{DD}. The write delay, on the other hand, is measured as the interval between the application of data at the bit-lines and the moment when the voltage difference between Q and QB reaches half of V_{DD}.

Monte Carlo simulations were conducted to analyze the statistical distribution of read and write delays at three different temperatures: -25°C, 27°C, and 125°C. Figure 3.6 illustrates the histogram distributions for read, write-1, and write-0 operations. The mean delay values at 27°C for read, write-1, and write-0 are 365.226 ps, 89.57 ps, and 29.13 ps, respectively. The results indicate that temperature variations significantly influence delay, with higher temperatures

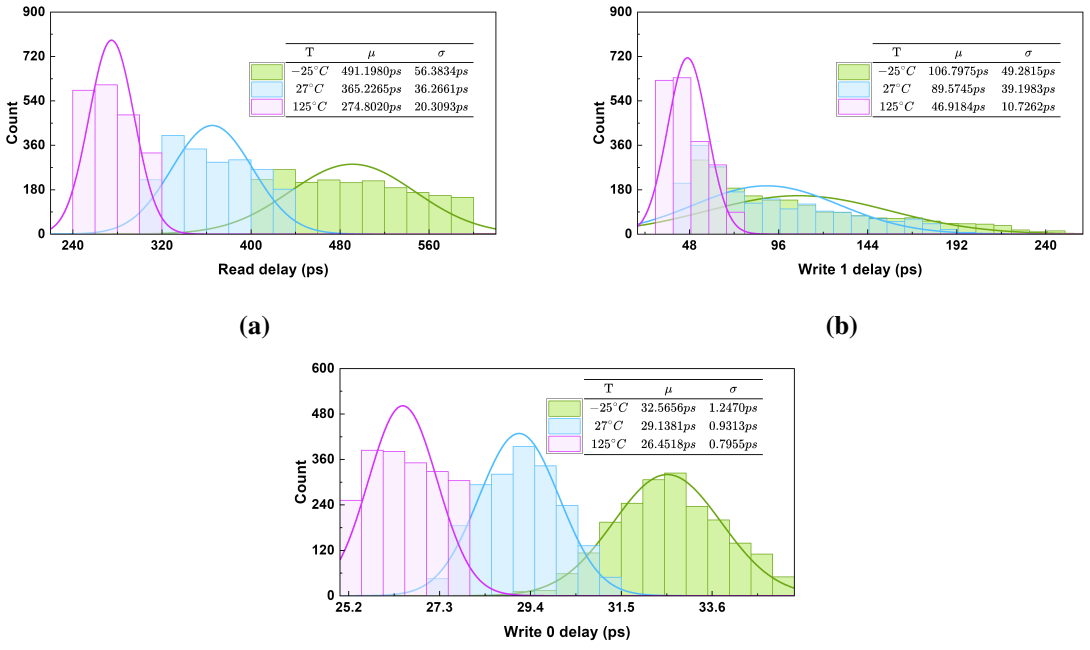


Figure 3.6: Temperature-dependent delay analysis of the proposed 8TS SRAM bit-cell for read and write operations at V_{DD} = 0.8 V. (A) Read delay, (B) Write-1 delay, and (C) Write-0 delay histograms, highlighting the impact of temperature variations on SRAM cell performance.

Table 3.2: Coefficient of Variation (CoV) for Read and Write Delays at Different Temperatures

Operation	-25°C CoV (%)	27°C CoV (%)	125°C CoV (%)
Read Delay	11.48	9.93	7.39
Write-1 Delay	46.14	43.76	22.86
Write-0 Delay	3.83	3.20	3.01

leading to reduced delay values, while lower temperatures cause an increase. A key observation is the variation in delay values across different temperatures. When the temperature is reduced from 27°C to -25°C, the delay increases by 1.3449, 1.1922, and 1.1176 times for read, write-1, and write-0 operations, respectively. Conversely, increasing the temperature to 125°C results in delay reductions of approximately by 24.76%, 47.63%, and 9.22% for read, write-1, and write-0 operations, respectively. This trend aligns with the expected behavior of semiconductor devices, where higher temperatures enhance carrier mobility, thereby reducing delay.

To quantify the impact of process variations, the coefficient of variation (CoV) is calculated for each delay type at different temperatures. Table 3.2 presents these values, computed as $\frac{\sigma}{\mu} \times 100\%$. The results reveal that write-1 delay exhibits the highest variability, particularly at lower temperatures, whereas write-0 delay remains relatively stable across all conditions. These results demonstrate that the proposed 8TS SRAM cell maintains acceptable delay variation across a wide range of temperatures. The observed trends indicate that the cell is well-suited for applications requiring stable performance under varying environmental conditions.

3.3.5 HALF-SELECT ERROR RESILIENCE OF 8TS SRAM CELL

Static Random-Access Memory (SRAM) is a fundamental component in high-speed cache and embedded memory architectures, where reliable data retention is crucial for overall system stability. One of the key challenges in SRAM design is mitigating the half-select problem, which arises due to shared control signals among adjacent bit-cells. This issue can lead to unintentional disturbances in non-targeted cells, potentially flipping stored bits and causing data corruption. The proposed 8TS SRAM bit-cell is designed to effectively eliminate such disturbances and ensure reliable operation under various read and write conditions. To validate its resilience, a 2×2 array of 8TS SRAM cells is constructed, as shown in Figure 3.7(A), where the selected, half-selected, and unselected cells are clearly highlighted.

The 2×2 SRAM array consists of four bit-cells: CELL 0, CELL 1, CELL 2, and an unselected cell (CELL 3). These bit-cells share control signals such as word lines (WL) and read bit lines (RBL), which dictate memory access and operation. Specifically, W1 is shared between CELL 0 and CELL 1, while W2 is shared between CELL 2 and the unselected cell. Similarly, the read bit line RBL1 connects CELL 0 and CELL 2, while RBL2 is shared between CELL 1 and the unselected cell (CELL 3). Since multiple cells depend on these shared signals, a write operation targeting one cell can partially activate its neighboring cells, making them susceptible to disturbances. If the stored data in these half-selected cells is unintentionally flipped due to an

operation performed on the selected cell, this results in a half-select disturbance.

A robust SRAM design must prevent such disturbances to maintain data integrity. To analyze the behavior of the proposed 8TS SRAM bit-cell under these conditions, transient simulations were conducted, and the results are depicted in Figure 3.7(B). The first subplot illustrates the applied control signals over time, indicating that WL1 and WBL1 are activated to perform a write operation on CELL 0. The second subplot captures the transient response of CELL 0, the selected bit-cell, showing a successful write operation where storage node Q0 transitions from high to low while its complementary node QB0 switches from low to high. The third subplot presents the response of CELL 1, a half-selected cell that shares W1 with the selected cell. Despite this shared wordline, the storage nodes Q1 and QB1 maintain their original states, confirming that CELL 1 remains unaffected by the operation on CELL 0. Similarly, the fourth subplot depicts the behavior of CELL 2, another half-selected cell due to its connection to the shared read bit line RBL1. The data stored in CELL 2 remains unchanged, demonstrating that the half-select resilience extends across both wordline and bitline-shared cells. Lastly, the unselected cell remains in hold state throughout this operation since none of its control signals are activated, providing additional validation that the proposed 8TS SRAM bit-cell effectively prevents half-select disturbances and ensures reliable data retention across the array.

The results confirm that the proposed design successfully mitigates half-select disturbances, preserving data integrity in both half-selected and unselected cells. A clear indication of half-select resilient cells is the inclusion of row and column exclusive control signals which yield the design soft-error free. Even during aggressive write operations, the stability of storage nodes in neighboring cells remains intact, preventing unintended bit flips and improving the overall robustness of the SRAM array. This level of resilience is particularly important for high-speed and low-power applications where minimizing error rates is critical. By addressing the half-select

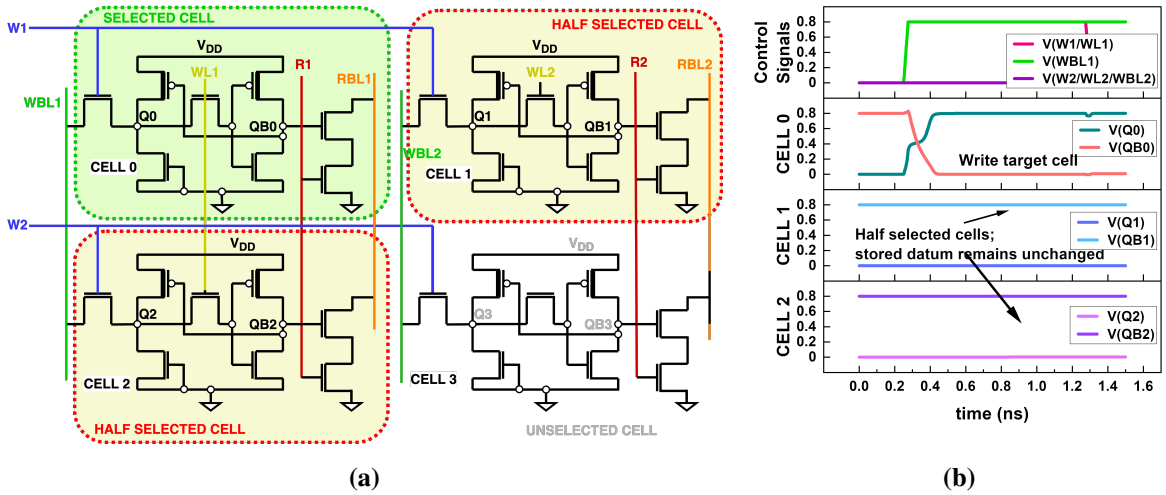


Figure 3.7: Illustration of half-select resilience in the proposed 8TS SRAM bit-cell. (A) A 2×2 array of 8TS SRAM cells, highlighting the selected, half-selected, and unselected cells. (B) Transient response of control signals and storage node voltages for selected, half-selected, and unselected cells, demonstrating the robustness of the proposed design.

problem at the circuit level, the proposed 8TS SRAM bit-cell enhances reliability, making it ideal for next-generation memory architectures that require high stability and efficiency.

3.4 COMPARATIVE ANALYSIS OF 8TS SRAM CELL WITH OTHER EXISTING SRAM ARCHITECTURES

Various SRAM bit-cells with different transistor counts have been proposed by researchers over the past few decades, each offering unique trade-offs in terms of stability, power consumption, and performance. The single-ended dual-port 7T SRAM cell [168], shown in Figure 3.8(A) incorporates three control signals: WWL, RWL, and FCS. During the read operation, RWL is activated while WWL and FCS remain off, enabling dynamic read decoupling to minimize read disturbance. For a write-0 operation, all three control signals are high, ensuring a strong write operation. Conversely, in the write-1 operation, only WWL is high, while RWL and FCS remain low. This configuration leverages dynamic feedback cutting, effectively reducing power consumption compared to the conventional 6T SRAM cell.

Fig. 3.8(B) illustrates a positive feedback controlled eight transistor (8Ta) SRAM cell [169]. This cell has single port, differential-ended topology whose write is differential-ended

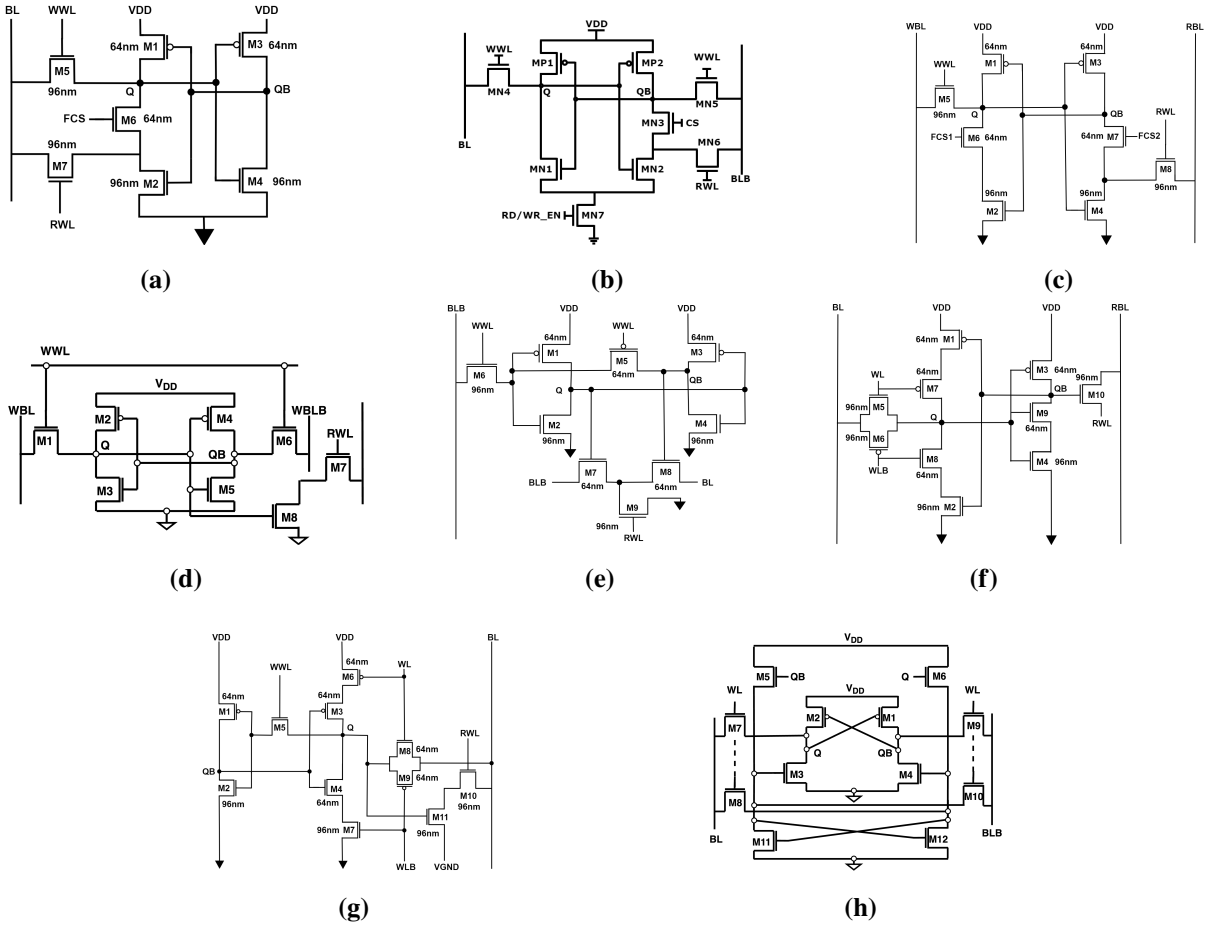


Figure 3.8: Circuit schematics of existing SRAM bit-cells with varied transistor count (A)7T, (B) 8Ta, (C) 8Tb, (D)8Tc, (E) 9T, (F) 10T, (G) 11T, and (H) 12T.

operation while read operation is of single-ended nature. The cell uses PMOS MP1, MP2 and NMOS MN1, MN2, MN3 transistors to form the inverter latch. Write operation is governed by the WWL wordline signal, connected to the gate of MN4, MN5 access transistors whereas read operation is initiated by RWL read wordline signal. Transistor MN3, controlled by CS control signal, disconnects feedback path between V_{DD} and GND to facilitate read SNM free operation. Transistor MN7, controlled by RD/WR_EN control signal, is common to whole row of memory array and utilized to facilitate reduced leakage power for hold operation. However, cell experiences high dynamic power dissipation and higher read latency owing to the utilization of differential write structure and three series connected NMOS transistor in the read path respectively.

The 8Tb single-ended SRAM cell with dynamic feedback control, presented in Figure 3.8(C), was first presented in [170]. This architecture separates the read and write operations using distinct bit-lines, RBL and WBL, controlled by feedback signals FCS1 and FCS2. The cell achieves read decoupling by isolating the read path from the write operation, which enhances its robustness. This structural modification significantly improves data stability and static noise margin (SNM), making it particularly effective in sub-threshold operation, where conventional 6T SRAM cells often suffer from read instability and increased leakage.

The conventional 8Tc cell, shown in Figure 3.8(D), was proposed as an improved alternative to the traditional 6T SRAM structure [171]. This design introduces two additional transistors to decouple the read and write operations, thereby eliminating read disturbances that typically degrade stability in a standard 6T cell. By incorporating a dedicated read path, the 8Tc cell prevents the pass-gate transistor from pulling up the storage node during a read operation, which significantly enhances static noise margin (SNM) and enables robust operation even at lower supply voltages. This cell features separate read and write word lines, allowing independent access to stored data without affecting the internal storage nodes. The dedicated read mechanism ensures that the worst-case stability condition in the 8Tc configuration is reduced to that of two cross-coupled inverters, which inherently provides a much larger SNM compared to 6T SRAM. Furthermore, despite requiring an additional 30% area overhead compared to a minimum sized 6T cell, the 8T structure facilitates continued technology scaling while maintaining improved stability, particularly under aggressive voltage scaling.

The 9T SRAM cell, depicted in Figure 3.8(E), retains the fundamental 6T configuration [172], with the addition of a feedback transistor (M5) that activates during the write operation. This modification effectively reduces the activity factor, lowering dynamic power consumption during write cycles [173]. Additionally, the breaking of the feedback loop between the cross-coupled inverters enhances write ability, enabling faster write operations while maintaining stability. The single-ended nature of the 9T design further contributes to reduced write power consumption, making it an energy-efficient alternative for low-power SRAM applications.

The 10T SRAM cell, illustrated in Figure 3.8(F), was developed by Eslami et al. [174]. This cell features a dual-port architecture, where a dedicated read port is formed by transistor M10, RBL, and RWL, allowing non-destructive read operations. The write port, in contrast, consists

of a PMOS-NMOS transmission gate (M6 and M5) directly connected to the bit-line (BL). By utilizing this separate read path, the cell ensures improved read stability while reducing leakage power. Additionally, the enhanced transistor arrangement provides significant improvements in read, write, and hold stability, particularly in sub-threshold operation, making it a viable choice for ultra-low-power applications.

The 11T SRAM cell, as described in [175] and represented in Figure 3.8(G), introduces further architectural refinements to enhance read and write stability. The inclusion of a feedback transistor (M5) selectively connects or isolates the inverters during different operations. Specifically, during the write operation, M5 is activated, facilitating efficient state switching. Conversely, in the read and hold modes, the two inverter pairs, consisting of transistors M1-M2 and M3-M4, remain disconnected by M5, preventing unintended disturbances. This design significantly enhances both read and write SNM while exhibiting reduced read and write delays compared to conventional 6T SRAM cells. The additional transistors contribute to improved robustness, particularly under low-voltage operation, making it well-suited for energy-constrained environments.

The 12T SRAM cell, shown in Figure 3.8(H), was introduced by Zhao et al. [176] as a radiation hardened 12T memory bit-cell, specifically optimized for space applications where resilience to single-event upsets (SEUs) is critical. Unlike conventional designs, this cell employs 10 NMOS transistors and only 2 PMOS transistors, leveraging polarity reversal mechanisms to mitigate charge-sharing effects caused by radiation strikes. The cell integrates four storage nodes (Q, QB, S0, and S1) to store complementary data, allowing rapid state recovery after an SEU event. Additionally, the 12T cell features enhanced write-ability and reduced write delay, significantly improving performance in high-frequency and low-voltage environments.

The delay values for read and write operations across various 8T SRAM bit-cells are presented in Figure 3.9(A). The 8TS cell exhibits a significantly lower read delay of 360.93 ps compared to 8Ta, 8Tb, and 8Tc, which have read delays of 479.32 ps, 480.85 ps, and 481.11 ps, respectively. Similarly, the write-0 delay of the 8TS cell is lower then 8Ta cell, while 8Tb and 8Tc offer slightly better write-0 delay. The write-1 delay follows a similar trend, where the 8TS cell achieves 76.99 ps, whereas 8Tc demonstrates the highest write-1 delay at 80.66 ps. These results indicate that the 8TS cell maintains competitive performance in read and write operations while ensuring lowest read delay across all three operations.

Figure 3.9(B) presents the power consumption characteristics of the analyzed 8T SRAM bit-cells. The 8TS cell demonstrates one of the lowest read-1 and write-1 power consumptions, highlighting its energy efficiency. The write-0 and read-0 power values are slightly higher than those of certain counterparts; however, this is compensated by the relatively lower delay exhibited by the 8TS cell. Additionally, the power trends suggest that while 8Ta and 8Tb cells exhibit higher write power consumption, the 8TS cell maintains an optimal balance between power efficiency and performance. The write margin (WM) values, including WM-1 and WM-0, for different 8T SRAM cells are illustrated in Figure 3.9 (c). The 8TS cell exhibits a normalized WM-1 of 1.288 and normalized WM-0 of 1.346. While the 8Tc cell demonstrates the lowest WM-0 at 0.44226 V, the 8Ta and 8Tb cells show relatively higher write margins. The observed

variations stem from the different transistor arrangements and control mechanisms employed in each bit-cell architecture.

The hold static noise margin (HSNM) and read static noise margin (RSNM) values for the analyzed 8T SRAM cells are depicted in Figure 3.9 (d). The 8TS cell achieves a HSNM/RSNM of 0.23242 V at V_{DD} = 0.8V. The stability of the 8TS cell remains moderate, ensuring reliable read operations under varying voltage conditions. These results collectively indicate that the 8TS SRAM bit-cell offers a balanced trade-off between read/write delay, power consumption, and noise margin, outperforming other 8T configurations in multiple key metrics.

The delay characteristics of SRAM architectures classified by transistor count, including the proposed 8TS cell, are illustrated in Figure 3.10 (a). The 8TS cell achieves a balanced read and write delay, demonstrating lower write-1 delay than the 7T, 9T, 10T, 11T, and 12T cells. The read delay of the 8TS cell is significantly lower than that of the higher transistor count designs,

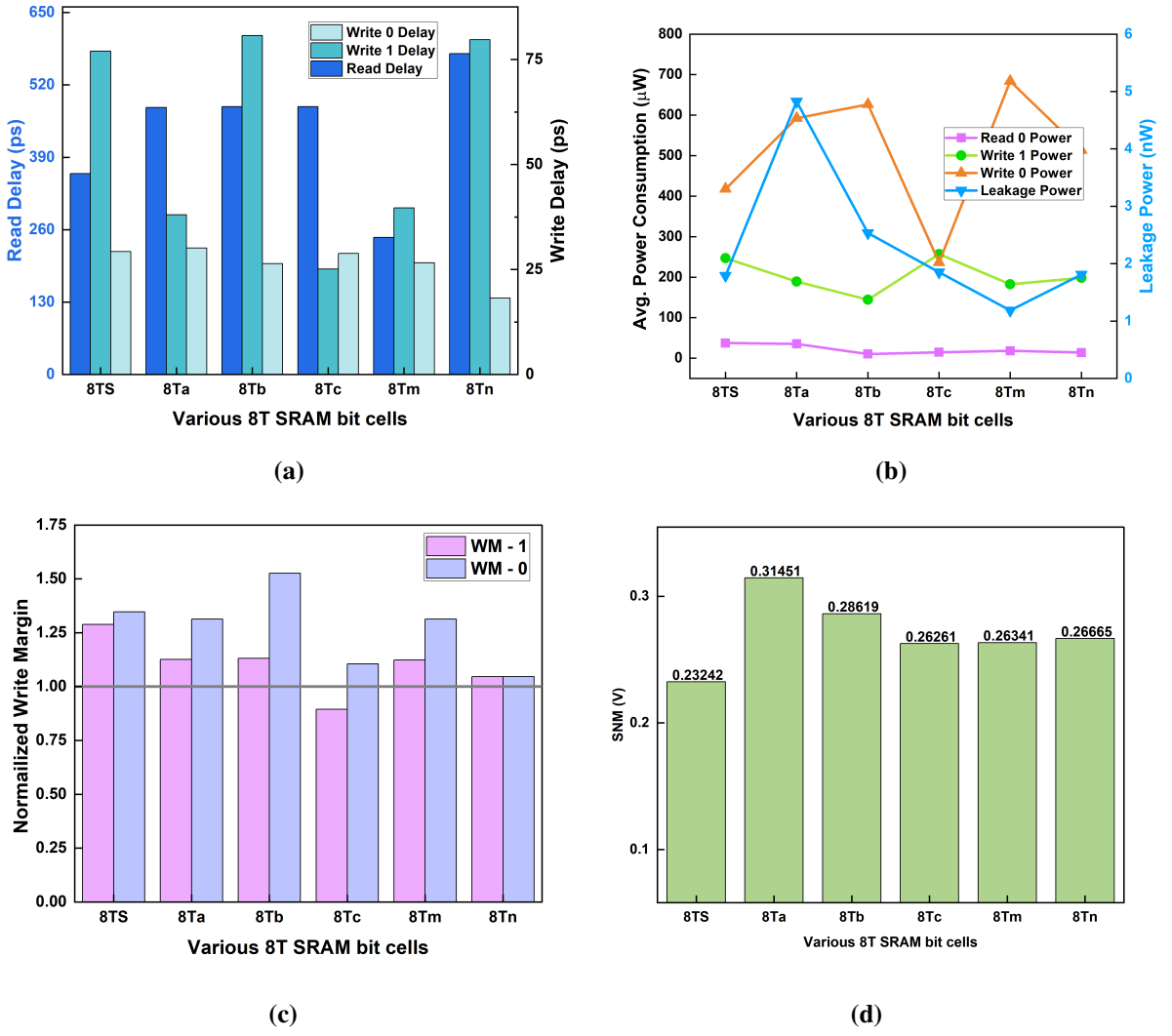


Figure 3.9: Performance comparison of the proposed 8TS SRAM bit-cell with existing 8T SRAM variants: (A) Read/Write delay, (B) Power consumption, (C) Normalized Write Margin (WM) and (D) Static Noise Margin (SNM).

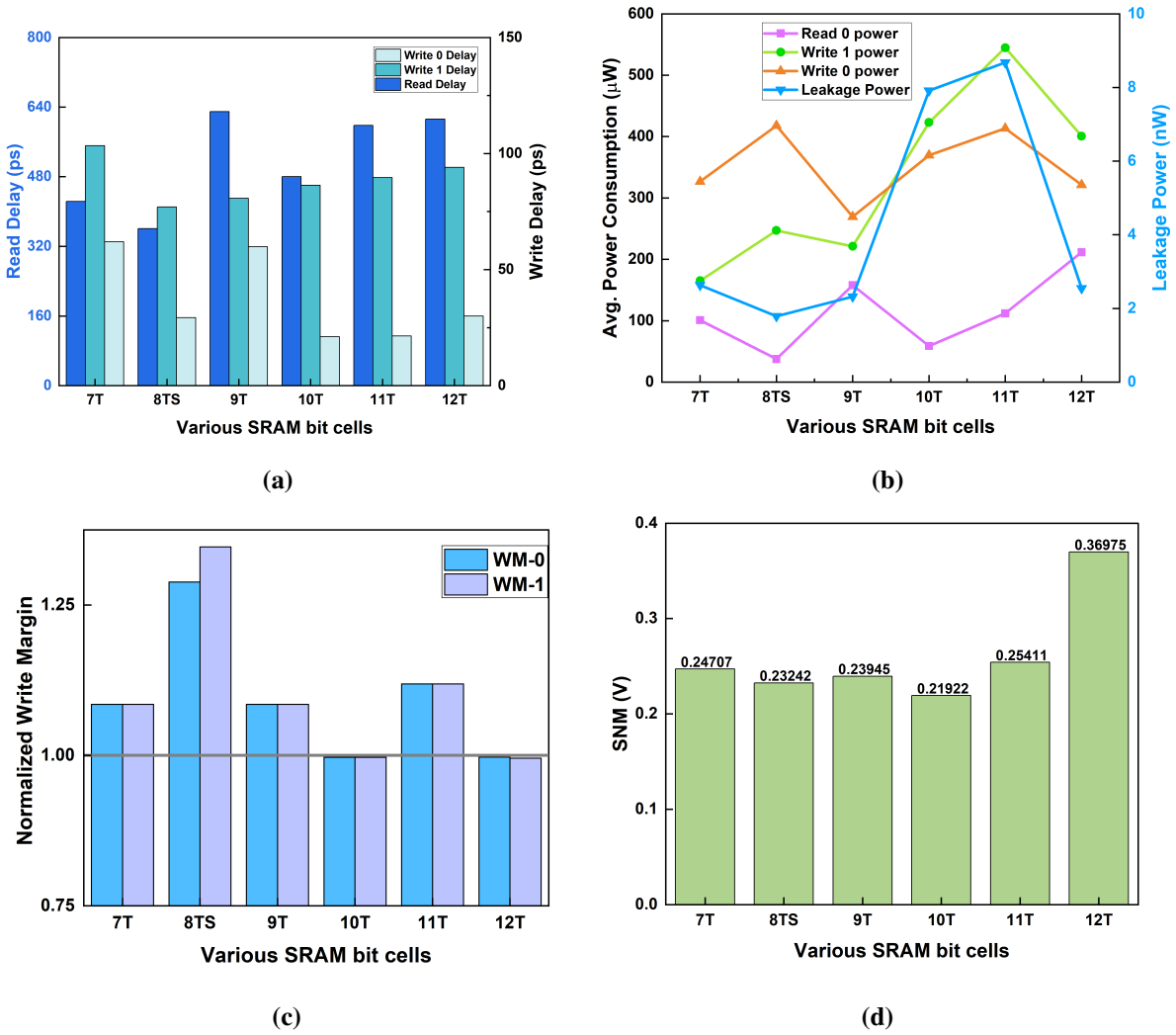


Figure 3.10: Comparative analysis of the proposed 8TS SRAM bit-cell compared to existing SRAM bit-cells with varying transistor counts: (A) Read/Write delay, (B) Power consumption,(C) Normalized Write Margin (WM) and (D) Static Noise Margin (SNM).

ensuring improved read access time. The write-0 and write-1 delays of the 8TS cell are also well-optimized, making it competitive with traditional 7T and 8T designs while outperforming the larger transistor count SRAM variants in terms of speed.

Figure 3.10 (b) presents the power consumption characteristics for various SRAM bit cells. The proposed 8TS cell maintains a favorable balance between read and write power. Notably, the read-0 and write-1 power consumption of the 8TS cell are among the lowest in comparison to other cells, signifying its energy efficiency. Although the write-0 power is slightly higher, it remains within an acceptable range when considering the trade-off with reduced delay. The higher transistor count designs, such as 10T, 11T, and 12T, exhibit significantly increased power consumption, making the 8TS cell a more power-efficient alternative. The write margin (WM) values, normalized for various SRAM bit cells, are depicted in Figure 3.10 (c). The 8TS cell exhibits one of the highest WM-0 and WM-1 values among all compared designs, indicating reduced write stability. Compared to traditional 8T and higher transistor count cells, the 8TS

cell achieves enhanced write robustness, ensuring reliable data storage under different operating conditions.

The Static Noise Margin (SNM) values of different SRAM cells are compared in Figure 3.10(d). The proposed 8TS cell achieves a hold SNM (HSNM) of 0.23242 V, which is comparable to other high-performance SRAM cells. While the 12T cell attains the highest SNM value of 0.36975 V, it comes at the cost of significantly increased power consumption and area overhead. The 8TS cell aims to strike an optimal balance by providing a competitive SNM while maintaining lower power and delay characteristics, outperforming many conventional SRAM architectures in overall efficiency.

The area overheads of different SRAM bit cells compared to the baseline 7T cell are shown

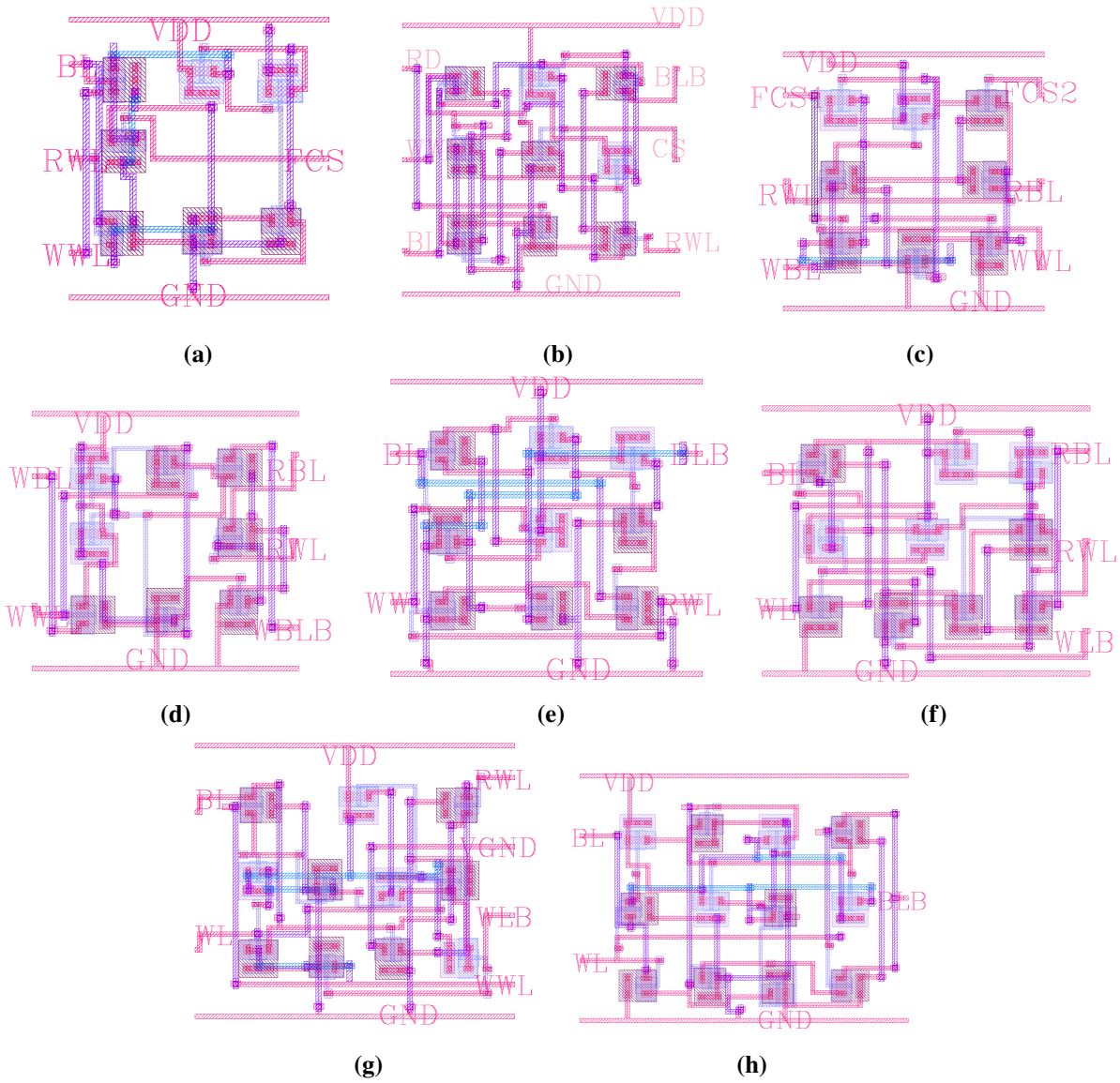


Figure 3.11: Layout representations of different SRAM bit-cells (7T–12T) used for comparative analysis, illustrating transistor arrangement and routing complexity: (A) 7T, (B) 8Ta, (C) 8Tb, (D) 8Tc, (E) 9T, (F) 10T, (G) 11T, and (H) 12T. Note: Layouts are not optimized for array-level implementation.

Table 3.3: Area overhead comparison of various SRAM bitcells relative to the 7T cell.

SRAM Cell	Area Overhead over 7T (%)
8Ta	42.17
8Tb	4.85
8Tc	2.95
8TS	4.86
9T	33.69
10T	31.64
11T	54.67
12T	70.61

in Table 3.3. Among the 8T variants, the 8Ta cell exhibits the highest area overhead at 42.17%, this is due the NMOS footswtich being included in the area calculations of the left most column of the array. This NMOS footswitch is shared across all the rows of the array thereby reducing the area footprint of all other cells in a row except the leftmost cell. The 8Tb, 8Tc, and 8TS cells maintain a minimal area increase of 4.85%, 2.95%, and 4.86%, respectively. The 9T and 10T cells incur a more significant area penalty of 33.69% and 31.64%, respectively, due to additional transistors required for improved stability. Higher transistor count cells, such as 11T and 12T, show the most substantial area overheads of 54.67% and 70.61%, respectively.

The layout structures of these bit cells, illustrated in Figure 3.11, visually highlight the increasing area complexity as transistor count increases. The 7T and 8T-based designs (Figures 3.11A-D) maintain relatively compact layouts, while the 9T, 10T, 11T, and 12T cells (Figures 3.11E-H) demonstrate increased diffusion area, interconnect complexity, and routing congestion. Despite the 8TS cell's slight area increase (4.86% over 7T), its enhanced performance, power efficiency, and stability make a compelling case for it's usage in low-power SRAM for voltage scaling.

3.5 PROPOSED DIFFERENTIAL 8T SRAM CELL ARCHITECTURE

The differential 8-transistor (8TD) SRAM cell architecture, shown in Figure 3.12, provides enhanced read stability, improved noise margins, and reduced power consumption compared to conventional 6T SRAM cells. This architecture employs a differential approach, separating the read and write paths, significantly minimizing the disturbance during read operations. The differential structure is composed of cross-coupled inverters (formed by transistors M1-M4) for data storage, along with dedicated read and write access transistors.

In this differential 8TD SRAM cell, transistors M2, M3, M4, and M5 constitute a bistable latch structure, ensuring stable storage of the complementary bit-lines (Q and Q_B). Transistors M1 and M6 serve as write-access transistors, enabling reliable data writing by connecting the internal storage nodes (Q and Q_B) to bitlines (BL and BLB) when the write word line

(WL) is asserted. Due to separate write access paths, the cell achieves robust write margins at scaled voltages. Transistors M7 and M8 function as feedback transistors controlled by the complementary write word line (WLB). These transistors significantly enhance the stability of the conventional 6T cell structure by reinforcing the stored data at the internal nodes (Q and Q_B). Additionally, by providing feedback during read operations, transistors M7 and M8 improve the cell's robustness against variations in process, voltage, and temperature (PVT).

3.6 SRAM STABILITY MARGIN COMPARISON WITH SINGLE-ENDED 8TS SRAM

Static Noise Margin (SNM) is a key metric used to evaluate the stability and robustness of SRAM (Static Random Access Memory) cells. It quantifies the maximum noise voltage that a memory cell can tolerate without altering its stored state. SNM is typically calculated using the butterfly curve, which is obtained by plotting the voltage transfer characteristics (VTCs) of the two cross-coupled inverters in the SRAM cell. To generate the butterfly curve, one inverter's VTC is plotted normally, while the other's is mirrored along the line y=x. The area of the largest square that can fit between the two curves—without crossing either—is considered the SNM. This square represents the maximum tolerable noise margin, and its side length gives the SNM value in volts. A higher SNM indicates a more stable SRAM cell that is less susceptible to noise-induced bit flips.

The butterfly curves for both the SRAM cells discussed are plotted at V_{DD} = 0.6 V in Figure 3.13 (a) and 3.13 (b) for the differential 8TD and the single-ended 8TS SRAM cells, respectively. These curves are derived by plotting the voltage transfer characteristics (VTCs) of the cross-coupled inverters and their mirrored counterparts, allowing the extraction of the Static Noise Margin (SNM) as the side of the largest possible square that fits between the two curves. At this supply voltage, the differential 8TD cell achieves an SNM of 0.19576 V, while the single-ended 8TS cell exhibits a lower SNM of 0.15308 V, highlighting the superior noise immunity of the

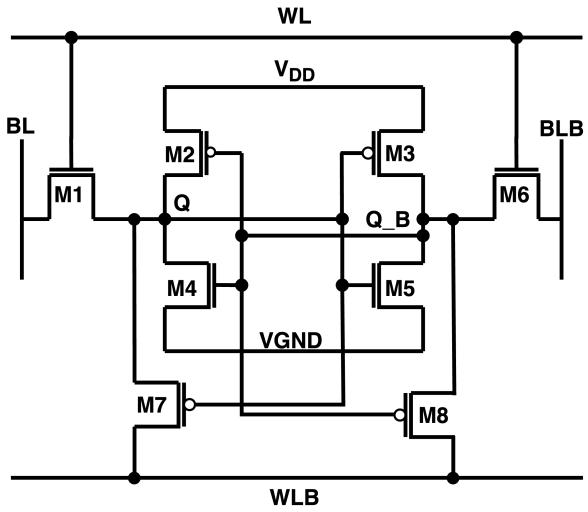


Figure 3.12: Circuit diagram of the proposed differential 8TD SRAM cell

differential structure under low voltage conditions.

However, this trend does not hold across all operating voltages. Figure 3.13 (c) explores the impact of supply voltage scaling on SNM. As V_{DD} increases, the SNM for both cells improves, but the 8TS cell begins to outperform the 8TD cell at higher voltages. This crossover in performance suggests that while differential designs are advantageous for ultra-low power or near-threshold operations, single-ended designs may be more favorable in regimes where power is less constrained and area or read-access simplicity is prioritized.

Furthermore, Figure 3.13 (d) illustrates the Write Margin (WM) of both 8TD and 8TS cells under two different write conditions; write 1 and write 0. It can be observed that write capability also scales with V_{DD}, and the single-ended 8TS design consistently offers higher write margins, especially under the condition labeled WM-1. This indicates that the 8TS architecture, although having a lower SNM at low voltages, provides improved writability, which may be crucial in write-intensive applications. Together, these plots offer a comprehensive picture of the trade-offs between stability and writability across voltage ranges in differential and single-ended SRAM architectures.

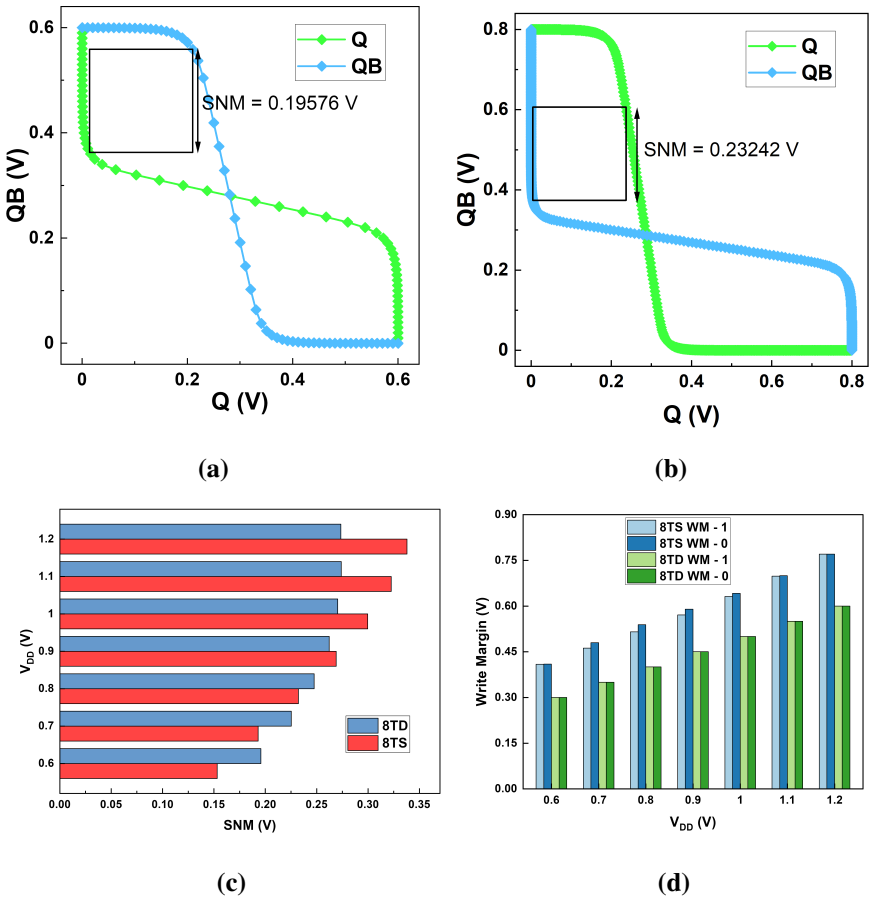


Figure 3.13: Butterfly Curve for calculation of static noise margin (SNM) (a) differential 8TD SRAM cell and (b) single-ended 8TS SRAM cell. (c) Impact of voltage scaling on (c) SNM and (d) Write margin (WM) of 8TD and 8TS cells.

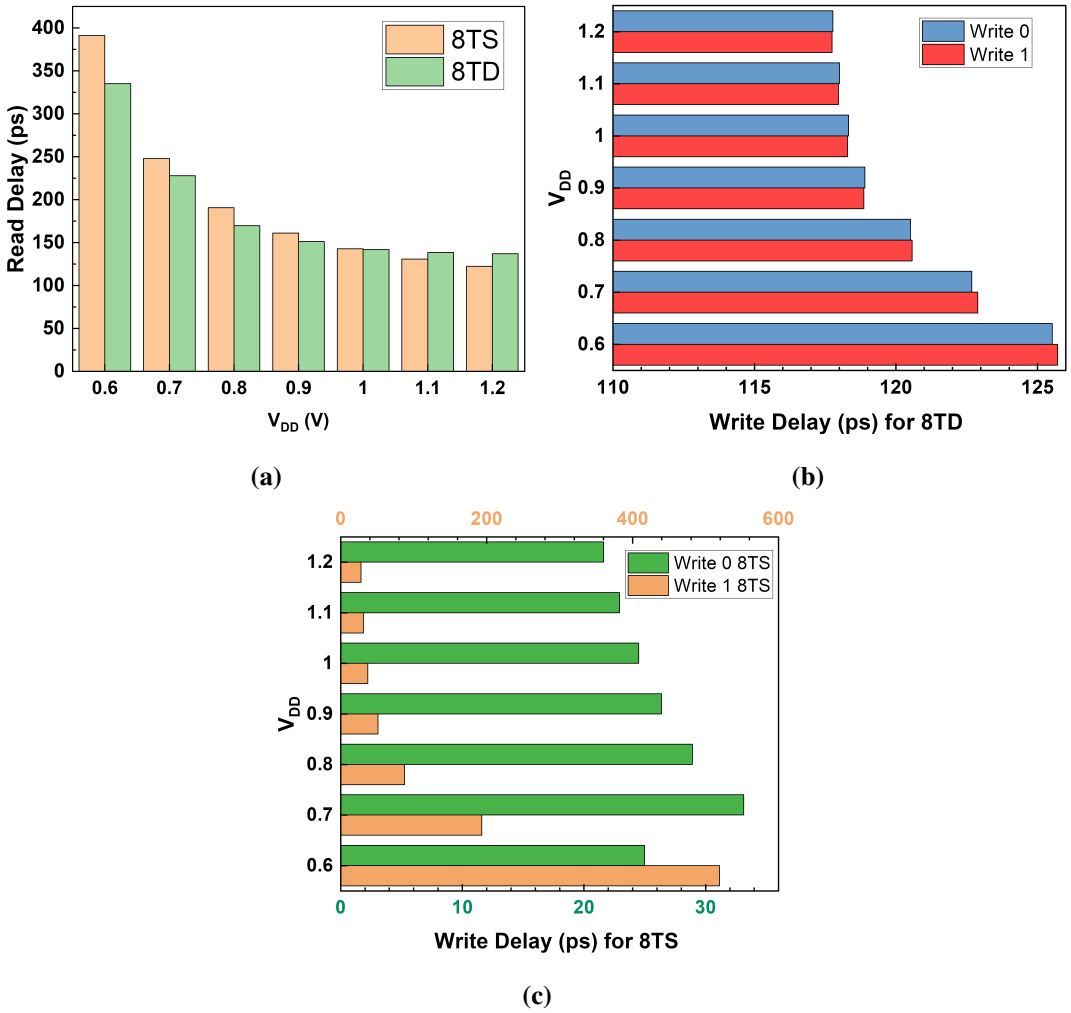


Figure 3.14: Timing analysis of the differential 8TD and single ended 8TS SRAM cells. (a) Read Delay comparison and Write Delay for (b) 8TD and (c) 8TS for V_{DD} = 0.6V to 1.2V

3.7 DELAY AND POWER ANALYSIS FOR READ AND WRITE OPERATIONS

In SRAM design, delay and power consumption are critical metrics used to evaluate the performance and efficiency of memory cells. Delay typically refers to the time required for a successful read or write operation and is usually quantified as read delay or write delay, depending on the operation under consideration. It is commonly measured from the time the wordline is activated to the point where the output either starts or completes a valid transition. Delay is influenced by factors such as bitline capacitance, cell drive strength, and sensing scheme, and is often characterized through transient simulations under varying supply voltages and loading conditions. Power consumption in SRAM is broadly classified into dynamic power and static (leakage) power. Dynamic power is consumed during switching events—such as charging and discharging of bitlines, activation of wordlines, and transitions in internal nodes—and is proportional to switching frequency, capacitance, and supply voltage. Static power, on the other hand, arises from leakage currents in the transistors even when the cell is idle and not switching.

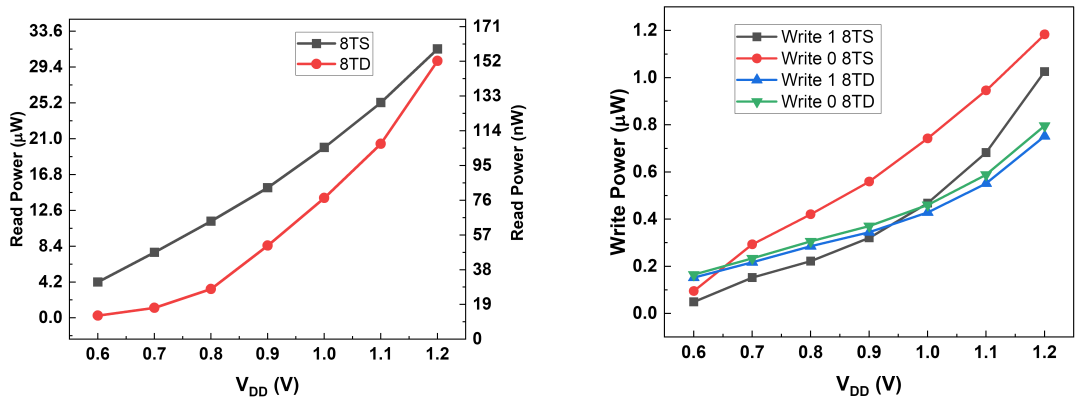


Figure 3.15: Comparison of (a) Read power (μW) (b) Write power (μW) of 8TS and 8TD SRAM Cells as a function of supply voltage

Power analysis typically involves measuring the energy consumed during read and write cycles, as well as evaluating leakage currents under different process corners and operating conditions. These metrics help guide design trade-offs between speed, energy efficiency, and robustness in SRAM architectures. Figure 3.14 presents a comparative analysis of read and write delays for the differential 8TD and single-ended 8TS SRAM cells over a range of supply voltages from $V_{DD} = 0.6V$ to $1.2V$. As shown in Figure 3.14(a), the 8TD cell consistently exhibits lower read delay compared to the 8TS cell across all voltage levels. At $V_{DD} = 0.6V$, the read delay of the 8TS cell exceeds 400 ps, whereas the 8TD cell achieves a significantly lower delay of around 340 ps. This gap narrows as V_{DD} increases, with both cells achieving comparable performance beyond $V_{DD} = 1V$, highlighting the superior low-voltage speed performance of the differential design.

Write delays for both cells are analyzed in Figures 3.14(b) and (c). For the 8TD cell [Figure 3.14(b)], write delays remain relatively stable across voltage scaling, ranging approximately between 110-125 ps, with a slight variation between writing '0' and writing '1'. In contrast, the 8TS cell [Figure 3.14(c)] demonstrates significantly longer write delays, particularly for writing a '0', which can exceed 30 ps at lower voltages. Writing a '1' is notably faster in the 8TS cell, especially at reduced V_{DD}, suggesting an asymmetry in write performance.

These observations indicate that the 8TD cell is more suitable for high-speed and low-voltage applications, offering consistent read and write performance. Meanwhile, the 8TS cell, although simpler and potentially more energy-efficient, shows limitations in both read speed and write symmetry, particularly under aggressive voltage scaling.

3.8 MONTE CARLO ANALYSIS FOR THRESHOLD VOLTAGE VARIATIONS

Monte Carlo analysis is a statistical simulation technique widely employed in VLSI design to evaluate the impact of process variations, particularly threshold voltage (V_{TH}) fluctuations, on circuit performance and reliability. In advanced nanoscale technologies, V_{TH} can vary significantly due to random dopant fluctuations, line-edge roughness, and other manufacturing

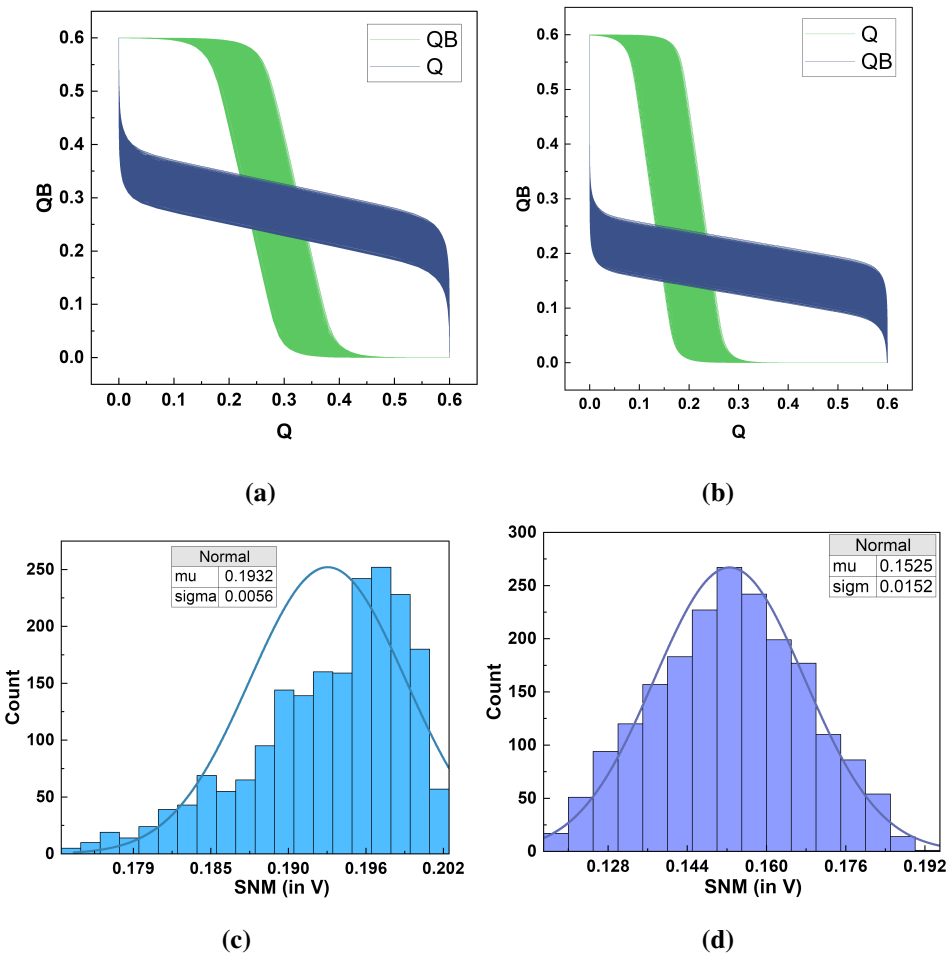


Figure 3.16: Monte Carlo (2000 point) analysis of differential 8TD and single-ended 8TS SRAM cells at $V_{DD} = 0.6V$. Butterfly curves for (a) 8TD (b) 8TS. The histograms for the SNM extracted from the butterfly curves for (c) 8TD and (d) 8TS.

imperfections. These variations lead to mismatches in transistor characteristics, which in turn affect key SRAM parameters such as static noise margin (SNM), read/write delay, and write margin. Monte Carlo simulations are carried out by executing a large number of circuit simulations, each incorporating randomly generated threshold voltage values drawn from a predefined statistical distribution, typically Gaussian. To assess the impact of process variation on the stability and performance of each cell, Monte Carlo simulations were conducted with 2000 runs, introducing threshold voltage (V_{TH}) variability modeled as a Gaussian distribution with $\sigma = 50mV$. This reflects realistic fabrication-level mismatches in advanced technology nodes. This approach yields a distribution of performance metrics instead of a single deterministic result, allowing designers to assess robustness, yield, and worst-case behavior. In SRAM analysis, Monte Carlo methods are crucial for capturing failures due to read disturb, write failure, or SNM degradation arising from asymmetry in inverter operation.

The outcomes of such simulations are often presented as histograms or cumulative distribution functions (CDFs) of critical parameters such as SNM, delay, or power. From these distributions,

statistical measures like mean, standard deviation, 3-sigma limits, and percentile-based yield estimates (e.g., 95% yield) can be derived. Figure 3.16 illustrates the results of a 2000-point Monte Carlo simulation performed to analyze the impact of threshold voltage variations on the Static Noise Margin (SNM) of the differential 8TD and single-ended 8TS SRAM cells at $V_{DD} = 0.6V$. The shaded butterfly curves shown in Figures 3.16(a) and 3.16(c) represent the spread of voltage transfer characteristics (VTCs) due to these variations, with each instance corresponding to a different threshold voltage sample. The 8TD cell [Figure 3.16(a)] demonstrates a tighter and more symmetric spread in its butterfly curves, indicating better resilience to process-induced mismatches compared to the 8TS cell [Figure 3.16(c)], which exhibits greater variability.

Figures 3.16(b) and 3.16(d) present the histograms of the SNM values extracted from the respective butterfly plots. The distribution for the 8TD cell shows a higher mean SNM of approximately $0.1932V$ with a relatively small standard deviation of $0.0056V$, suggesting consistent and robust noise immunity under threshold voltage perturbations. In contrast, the SNM distribution for the 8TS cell reveals a lower mean of $0.1525V$ and a noticeably larger standard deviation of $0.0152V$, pointing to reduced stability and a higher likelihood of noise-induced failures. These results confirm that the differential 8TD architecture is more tolerant to device-level variations, making it better suited for reliable operation in variation-sensitive environments.

3.9 SUMMARY OF IMPORTANT RESULTS

This chapter presented the design and analysis of SRAM bit-cell architectures aimed at improving stability, power efficiency, and robustness under low-voltage and variation-sensitive operating conditions. The proposed half-select-free single-ended 8TS SRAM bit-cell was evaluated with respect to static noise margin, write margin, read/write delay, dynamic power, leakage behavior, layout overhead, and Monte-Carlo variability. In addition, the performance of the differential 8TD cell was compared with the 8TS cell over a wide supply-voltage range to identify the operating regions where single-ended simplicity or differential robustness becomes more suitable. The analysis therefore provides a cell-level foundation for the later SRAM-IMC macro and array-level developments.

The principal findings of this chapter are outlined below–

1. A half-select-free, single-ended 8TS SRAM bit-cell has been designed in 32 nm CMOS. It achieves a static-noise-margin (SNM) of 0.232 V at $V_{\mathrm{DD}} = 0.8$ V—second only to a 12T reference cell—while limiting layout overhead to 4.86% relative to a 7T baseline.

2. Monte-Carlo sweeps (2000 samples across full PVT space) confirm 6σ robustness for SNM, write-margin ($\mathrm{WM}_1 = 0.515$ V), and read/write delay, validating the 8TS cell's tolerance to threshold-voltage mismatch, supply variation, and temperature fluctuation.

3. The 8TS cell records one of the lowest dynamic-power figures among all surveyed designs: 0.186 µW for read and 0.247 µW / 8.89 µW for write-1 / write-0, respectively, ensuring energy-efficient operation across the 0.5–1.0 V supply window.

4. The 8TS architecture is inherently half-select-error-free, enhancing reliability in large memory arrays without additional assist circuitry.

5. Two alternative eight-transistor cells were evaluated over 0.6–1.2 V: a differential 8TD cell and a single-ended 8TS cell. Near threshold 0.6–0.7 V the 8TD cell shows higher SNM and write-margin than the 8TS, providing better read/write reliability for low-voltage operation.

6. As V_{DD} rises toward 1.2 V, the stability gap between 8TD and 8TS narrows, yet the 8TD still offers faster access at the expense of extra control signals and a differential sense amplifier, which increases dynamic power.

7. Monte-Carlo analysis under threshold-voltage variation indicates that the 8TD retains a lower read-fail probability than the 8TS, further supporting its suitability for variation-tolerant, low-supply designs.

8. Overall, the 8TS cell provides the best compromise among stability, power, and area for single-ended arrays, whereas the 8TD cell is preferred when differential sensing and maximum low-voltage robustness are required; the 8TS cell remains a simpler, lower-power option at nominal voltages where stability margins are less critical whereas the 8TD cell trumps when high performance cell operation is key.

The results of this chapter show that no single SRAM bit-cell gives the best outcome for all operating requirements. The 8TS cell offers a compact and energy-aware solution for half-select-resilient single-ended SRAM arrays, with strong stability and low dynamic power under nominal and scaled supply conditions. The 8TD cell, on the other hand, provides improved low-voltage stability, faster access, and better tolerance to threshold-voltage variation because of its differential operation and compatibility with sense-amplifier-based readout. Thus, the chapter establishes a clear design trade-off between single-ended low-power operation and differential high-performance operation. This trade-off is important for the following chapters, where the proposed cell structures and sensing strategy are further extended toward SRAM-based in-memory computation.

CHAPTER – 4

8TD SRAM-IMC FOR BITWISE LOGIC IMPLEMENTATION

4.1 OVERVIEW OF THE CHAPTER

The explosive growth of data-centric applications—from large language models to real-time edge computing—has laid bare the energy and throughput limits of von-Neumann machines, where every arithmetic operation demands at least one costly *round-trip* across the memory–logic boundary. Recent studies show that data movement, not arithmetic, dominates system-on-chip (SoC) power in modern AI accelerators, often consuming $\geq 70\%$ of the total energy budget. *In-memory computing* (IMC) tackles this "memory wall" by performing computation within the storage array, thereby exploiting the memory's massive internal bandwidth. Among candidate media—STT-MRAM, ReRAM, PCRAM and others—CMOS static RAM (SRAM) stands out for its transistor-level maturity, sub-nanosecond access time, and seamless integration into existing digital flows. Yet, the canonical 6-transistor (6T) bit-cell couples read and write nodes, so naïvely activating multiple word lines to form Boolean functions risks destructive disturb, a limitation first analysed in the push-rule 6T SRAM-IMC macro. Over the last decade researchers have introduced an impressive taxonomy of SRAM topologies, peripheral circuits and algorithm–hardware co-design techniques that progressively enlarge the attainable compute set while safeguarding data integrity. In light of the preceding requirement, the objective is articulated as follows:

"Design and analysis of SRAM based in-memory computation (IMC) macro using modified sensing strategy"

Methodology

1. Develop an SRAM bit-cell array that embeds the modified sensing strategy required for in-memory computation (IMC).
2. Integrate the selected sensing topology into a complete IMC macro and verify correct arithmetic/logic functionality at the schematic level.
3. Evaluate the macro's energy, latency, and computational accuracy over a 0.5 V to 1.3 V supply range and for multiple word-lengths.
4. Perform process–voltage–temperature (PVT) corner sweeps and Monte-Carlo mismatch analysis to quantify robustness and yield.
5. Benchmark the proposed IMC macro against conventional SRAM + digital-logic baselines and published IMC designs, highlighting improvements in throughput, energy per operation, and area efficiency.

The chapter is organized as follows. Section 4.3 introduces the related work of SRAM IMC using standard 6T, 8T cells (differential and single ended) and illustrates the working principle and process of the circuit with examples. Section 4.4 analyzes how to optimize the 8TD cell proposed in Chapter 3 for IMC to perform logical operations and check the reliability of the cell in memory and compute modes. Section 4.5 presents the comparative results of all the discussed IMC structures.

4.2 INTRODUCTION

Technology scaling and the proliferation of AI-enabled edge and IoT platforms have drastically increased the amount of data that must be sensed, stored, and processed within tight power budgets [177]. At the same time, the historical gains obtained from CMOS scaling are no longer sufficient to deliver proportional improvements in energy efficiency and performance for strongly data-centric workloads. Although beyond-CMOS and emerging-device paradigms continue to attract attention as long-term alternatives, migrating the complete design ecosystem to new devices requires new fabrication processes, compact models, design methodologies, and verification flows, all of which introduce substantial cost and risk [178, 179]. Consequently, CMOS-compatible memory-centric acceleration remains one of the most practical near-term approaches for improving throughput without abandoning mature digital design infrastructures.

In the conventional von Neumann organization, memory and logic are physically separated, so even simple operations require repeated transfers of operands and results across the processor–memory interface. For modern workloads, especially those dominated by large volumes of feature maps, weights, and intermediate data, this communication overhead increasingly dominates both latency and energy. Compute-in-memory (CIM) addresses this limitation by either moving computation closer to the memory array, as in near-memory computing (NMC), or by embedding selected operations directly within the memory fabric, as in in-memory computing (IMC). A broad spectrum of memory technologies has therefore been investigated for CIM, including SRAM [180–183], RRAM [184, 185], DRAM [186, 187], and spin-transfer torque (STT) magnetic random access memory (MRAM) [188, 189]. RRAM- and MRAM-based solutions are attractive because they can provide non-volatility and high density, whereas DRAM-based schemes can leverage dense commodity-compatible arrays for data-parallel workloads. However, peripheral overhead, device non-idealities, and integration complexity strongly influence how efficiently such alternatives can be deployed in practical systems [184–186, 188]. In contrast, SRAM remains deeply embedded in modern processors, offers very low access latency, and maps naturally onto standard CMOS logic flows, making it especially attractive for low-latency on-chip IMC.

The IMC landscape can be broadly divided into analog and digital realizations. Analog IMC typically exploits charge sharing, current summation, or capacitive accumulation to perform vector–matrix or multiply–accumulate operations with high internal parallelism and attractive energy efficiency [190–192]. Its main limitation is that the final computational accuracy is closely tied to device mismatch, noise, reference generation, and, in many cases, the overhead

of analog-to-digital or mixed-signal interfaces. Digital IMC, by contrast, preserves discrete logic levels and therefore aligns more naturally with standard design and verification flows, while also offering better determinism for bitwise logic and low-precision arithmetic kernels [182, 193]. The key challenge in digital IMC is that correct functionality depends strongly on bitline discharge dynamics, simultaneous word-line activation behavior, and the robustness of the sensing strategy under process, voltage, and temperature (PVT) variation. As a result, the bit-cell structure and the peripheral read/compute circuitry must be co-optimized rather than designed in isolation.

This co-optimization is particularly evident in SRAM-based IMC. Early demonstrations such as X-SRAM showed that conventional 6T SRAM arrays can be repurposed to execute Boolean operations inside the memory array through suitable word-line control and sensing schemes [180]. However, the shared read/write path of the 6T cell makes concurrent word-line activation vulnerable to read disturb and constrains reliable compute operation. Decoupled-read 8T cells and related logic-in-memory SRAM structures address this problem by isolating the read path, improving noise margins, and enabling faster search or logic functionality with lower risk of corrupting the stored data [181, 194]. More recent digital SRAM-IMC macros have further extended the attainable compute set from simple Boolean primitives to higher-level arithmetic and in-array result-storage functions [182, 193]. Motivated by these developments, this chapter examines SRAM-IMC from the perspective of bitwise logic realization and emphasizes the role of differential sensing in achieving a favorable balance among reliability, latency, and energy efficiency.

4.3 IN-MEMORY COMPUTING USING STANDARD SRAM BIT CELLS

The conventional SRAM bit cell lends itself very well to IMC architectures due to the fact that two concurrently activated wordlines intrinsically perform NOR operation for the corresponding stored data values. This is clearly evident for standard 6T SRAM cell illustrated in Fig. 4.1 (a) with the sense amplifiers being used to perform the compute operation shown as their digital gate equivalent. The content addressable memory circuitry (CAM) implemented using standard 6T push rule cells [195] with the wordline split into two different control signals, WLA1 and WLA2 as shown in Fig. 4.1(a), performs XNOR logic operation to compare the search string with the stored values in the corresponding bit cells. This enables parallel search by applying the search string on the wordlines and comparing the column-wise stored values. Other iterations of 6T bit cell being used in IMC have also used sequential wordline activation instead of concurrently turning on multiple WLs which might cause read/ compute disturbance due to the common read/write path of conventional 6T cells. The cells in the same column will flip if the bitline discharges to a relatively low value, i.e., lower than the switching threshold of the single ended inverter based sense amplifiers and false read might occur. The 6T SRAM with under-driven multiple wordline control during the compute operation has also been used which helps avoid the read disturbance encountered when the stored datum 'A' and 'B' are different (AB = 01/10). However, the usage of low WL signals yields inherently slower speeds

for NMOS-access transistor based 6T. A solution to is to use a dual port 8T(6+2T) SRAM cell [181] with a isolated compute port comprised of two PMOS access transistors together with split WL which avoid the reliability degradation and increased latency suffered by 6T-based IMC architectures. The Fig. 4.3(a) illustrated the transient simulation for the two 6T cells accessed using sequentially activated WL signals and the corresponding discharging of the BL/BLB line based on the stored data combinations. The AND and NOR operations are implemented by using

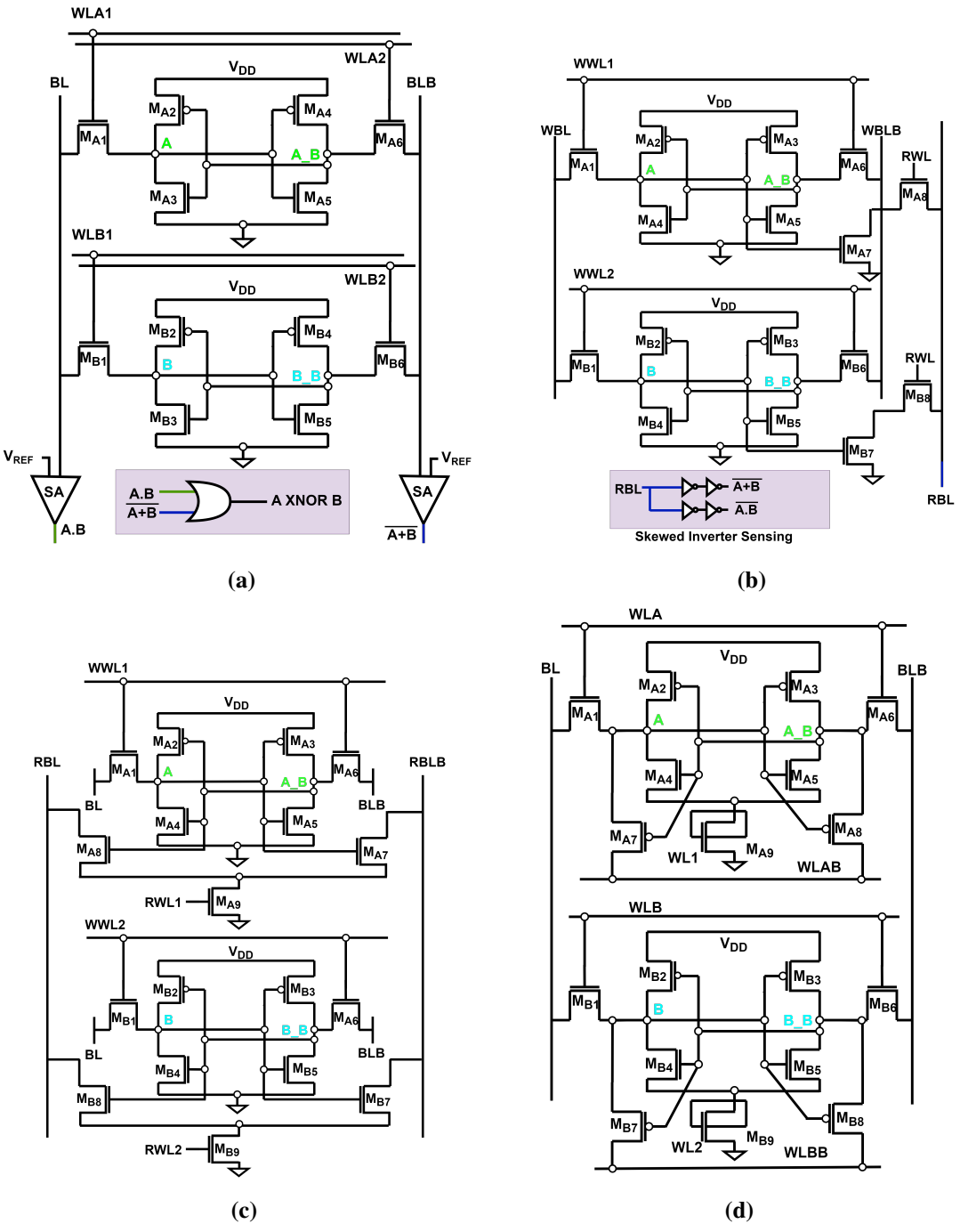


Figure 4.1: Circuit schematics for implementation of single bit boolean logic for two operands A,B along with the sensing strategy employed. (a) 6T SRAM-IMC (b) 8TC SRAM-IMC (c) 8TB SRAM-IMC and (d) 8TE SRAM-IMC

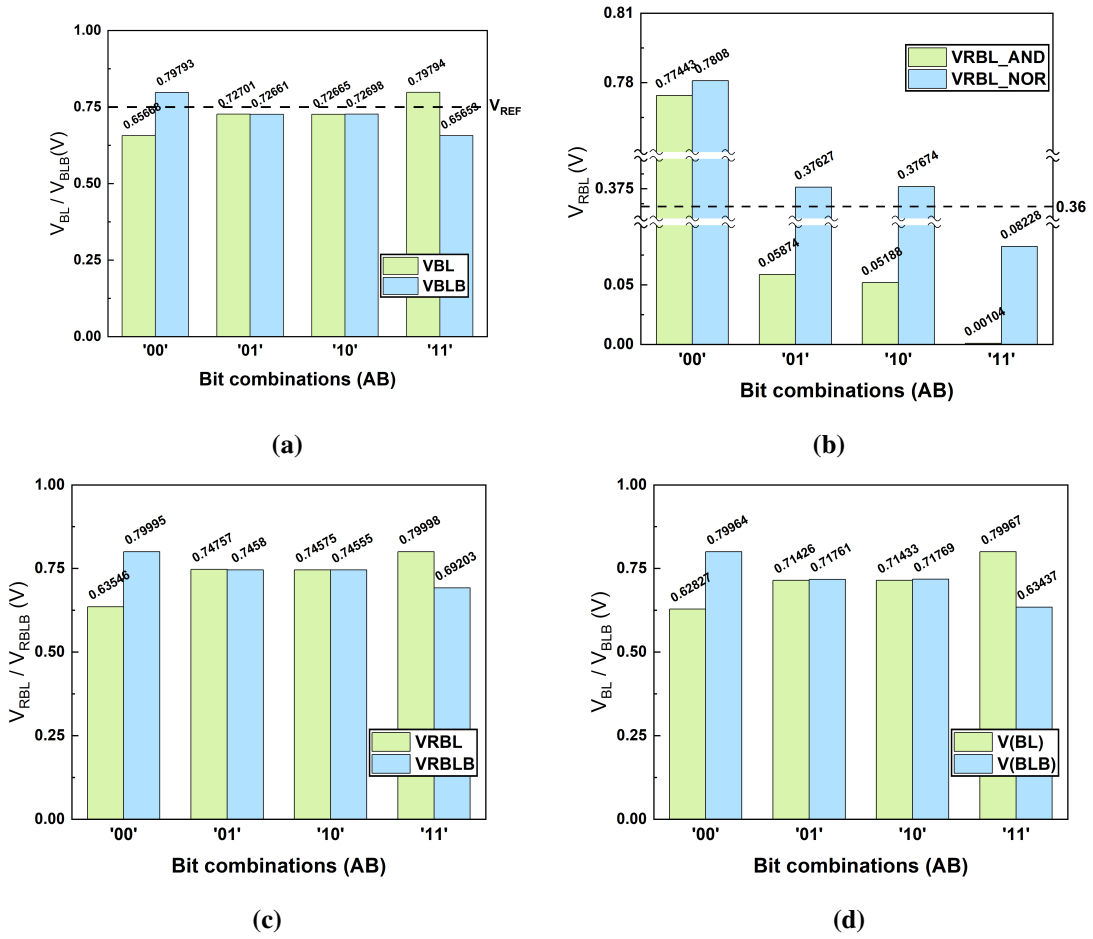


Figure 4.2: SRAM bit lines (BL/BLB) voltage levels to be sensed by SAs for IMC bitwise logical operations for (a) 6T SRAM cell (b) 8TC SRAM cell (c) 8TB SRAM cell and (d) 8TE SRAM cell

single ended SAs with V_{REF} taken such we get a high logic output for V_{BL} in case where AB = 11 for AND implementation and V_{BLB} where AB = 00 for NOR logic as shown in Fig. 4.2(a).

SRAM cell structures have evolved from the conventional 6T SRAM cell to single ended cell structures as well as differential ended modified architectures with increased transistor count. These are designed around an isolated read and write path which allows for modulation of the Static Noise Margin (SNM) in both the hold and read modes separately. The isolated read and write ports are also useful when multiple WL activations occur as is the case in Computation modes for standard SRAM cell structures. The read and write disturbance is mitigated using the isolated ports. The two conventional 8T (8TC) cells as shown in Fig. 4.1(b) store the corresponding operands for the two bit logic operations supported in-memory. The 8TC transient response for the logic operations supported is depicted in Fig. 4.3(b) where the Read bitline (RBL and RBLB) voltages are shown for all the possible bit combinations, i.e., 00,01,10,11. The charge and discharge of the RBL/RBLB is used as the output for implementing the logic operations. Skewed inverter sensing is used for performing the logic operation with the conventional cross-coupled inverter pair sharing the same bitlines acting as a wired NORed structure similar to the 6T SRAM-IMC. The resulting voltages and the outputs of the skewed

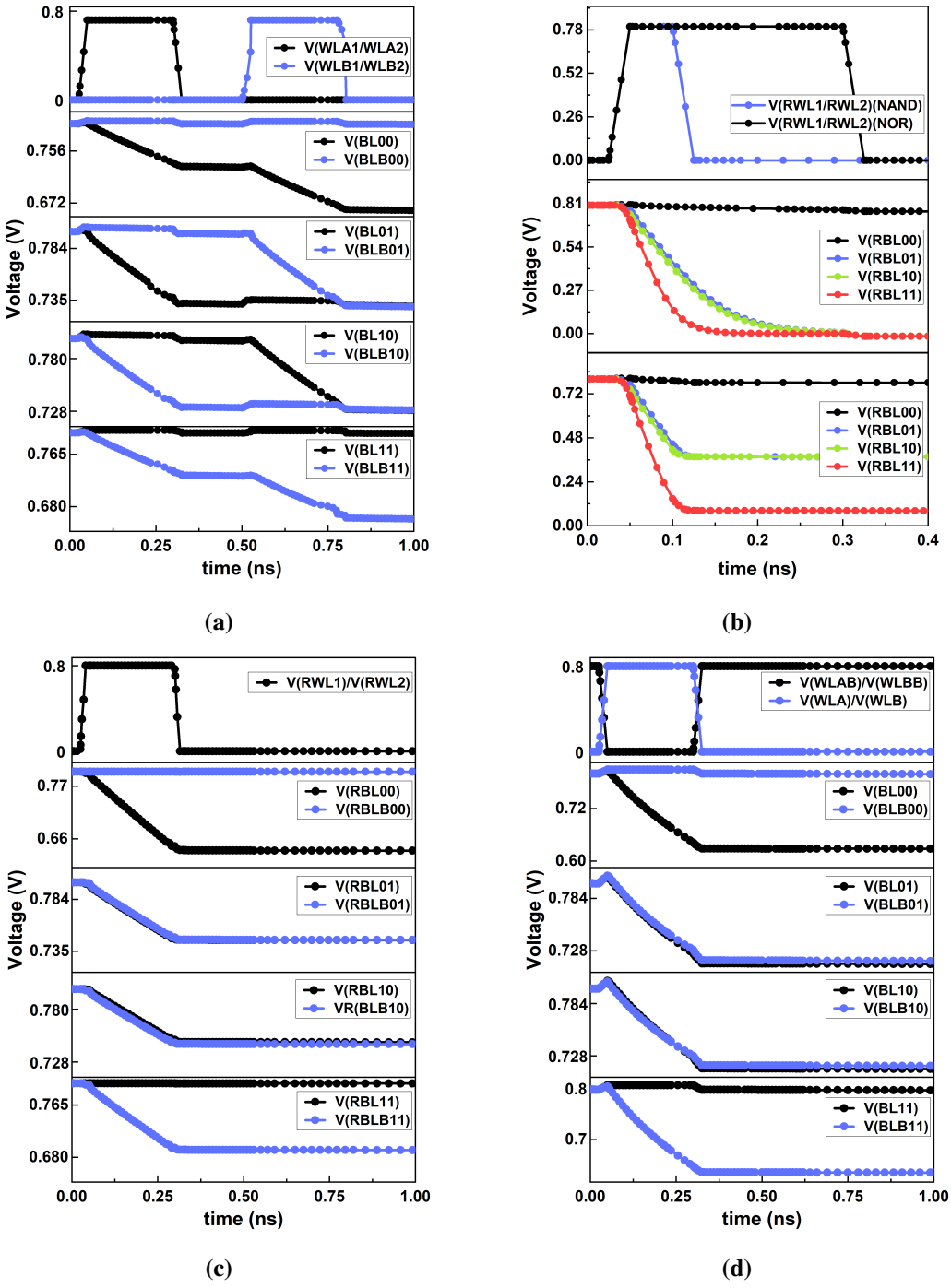


Figure 4.3: Transient response of SRAM bitcells in compute-mode for all possible bit combinations of two single bit operands (A,B) '00', '01', '10 and '11' at V_{DD} = 0.8 V for (a) 6T SRAM cell (b) 8TC SRAM cell (c) 8TB SRAM cell (d) 8TE SRAM cell

inverter sensing is shown in Fig. 4.2(b).

An asymmetric sense amplifier (SA) [180] is used with the two NMOS bitline access transistors (M8 and M9) sized non-uniformly based on the logic operation (NAND/NOR) to be realised. The differential 8T cell (8TB) [194] structure can be used in IMC for implementing logic operations like NAND, NOR and XOR. The cell's differential structure is designed using two

extra NMOS transistors together with a footswitch common to the whole row for WWL control signal connected in the pulldown path for the cross-coupled inverter pair. The circuit schematic for 8TB SRAM-IMC is shown in Fig. $4.1(c)$. The boolean logic functions are implemented using an asymmetric differential sense amplifier [180] which senses the differential Read bitlines (RBL/RBLB). The transistors are skewed by sizing the access transistors in the sense amplifier. The complement read bitline (RBLB) access transistor sized greater than RBL access transistor yield OR (NOR) logic combinations at the SAout (SAoutb) while a wider RBL access transistor results in AND (NAND) logic combination at the SAout (SAoutb). The RBL and RBLB voltages for different combinations of operands 'AB' are shown in Fig. $4.2(c)$. The transient response of the 8TB SRAM-IMC structure is depicted in Fig. $4.3(c)$ for input control signal WL/WWL at V_{DD} = 0.8 V.

The differential SRAM 8TE cell [196] and its adjacent structures lend themselves very well to in-memory computational needs due to their isolated read/write dual port architecture. A modified version of the cell is designed using the conventional cross-coupled inverter pair with two extra control buffer PMOS transistors being used together with a tail NMOS as shown in Fig. 4.1 (d). One of the control buffer transistors (Q = 1 connected to its gate terminal) forms a conducting path through the tail transistor during the access mode of the differential SRAM cell. While for the write operation to be performed the bitlines need a conduction path to the internal storage nodes the WLB = 0, and tail is turned ON.

4.4 IN-MEMORY COMPUTATIONS USING THE PROPOSED 8TD SRAM CELL AND MDSANF

In order to realize single-bit Boolean functions such as NAND/NOR and XOR/XNOR within the memory array, two identical instances of the proposed 8TD SRAM cell from Chapter 3 are connected in a column-wise configuration, as shown in Figure 4.4. The upper and lower dashed regions represent Cell 1 and Cell 2, respectively. These two cells store the input operands A and B, while their differential storage nodes provide both true and complementary data values to the shared bit-line/compute path. By selectively activating the corresponding control signals, the stored operands interact through the common differential bit-line structure, producing a logic-dependent voltage response. This two-cell arrangement therefore forms the proposed 8TD SRAM-IMC macro, where Boolean computation is performed directly using the stored data rather than transferring the operands to a separate logic unit.

There are two possibilities for the control wordline that lead to NAND/NOR or XOR/XNOR logic operations similar to the conventional 6T/8TB SRAM cells discussed in Section 4.3. When both the wordlines of the two cells storing the operands are activated at the same instant of time, i.e. concurrent wordline activation, the transient response shown in Fig. 4.5 (a) for the voltages at the bit line (BL) and complementary bit line (BLB) which is analogous to 6T SRAM-IMC. This is illustrated for all possible combinations of AB, namely '00,01,10,11'. Correspondingly, for sequentially activated wordlines of cell A followed by cell B, which helps mitigate the problem of read disturbance where the stored values might change when two or more wordlines in the

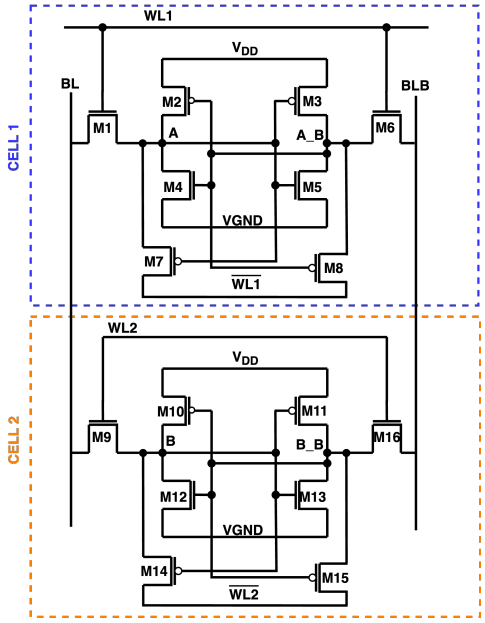
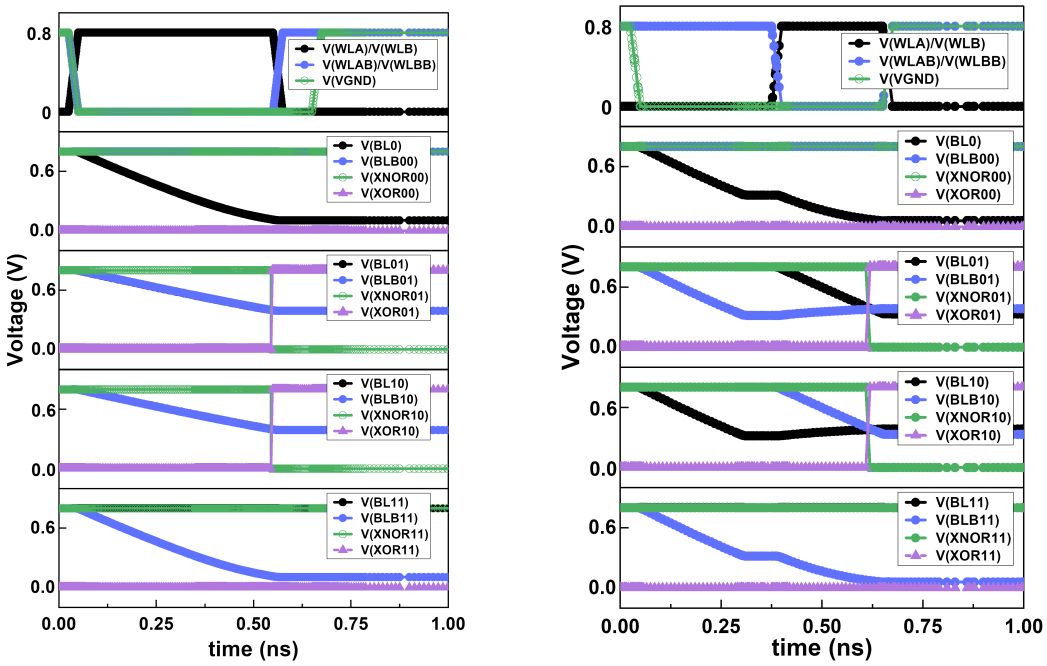


Figure 4.4: Circuit schematic for implementation of single bit boolean logic (NAND/NOR and XOR/XNOR) for two operands A,B using proposed 8TD SRAM cell

same column are turned on concurrently, the resulting transient simulation is depicted in Fig. 4.5 (b). The transient response obtained from the simulations demonstrates that the sensing mechanism produces accurate results, with no erroneous behaviour detected throughout the operation window. This ensures that the design yields zero computational errors, confirming its resilience against threshold variations in performing logic operations. This configuration effectively supports all possible input combinations for operands A and B, providing correct logical outputs without any errors.

To further validate the robustness of the design, Monte Carlo simulations for 6k trials were performed to showcase the reliability of the NAND/NOR and XOR/XNOR logic operations using the 8TD SRAM cell at T = 300.15 K for the TT corner with V_{DD} = 0.8 V and ΔV_{TH} = 50 mV to account for process variations. The results presented in Fig. 4.6 (a) and Fig. 4.6 (b) illustrate the transient output voltage for the MDSANF when performing NAND and NOR operations, respectively. These figures highlight the stability and consistency of the outputs across multiple simulation runs, further reinforcing the reliability of the 8TD SRAM-IMC macro. Figures 4.6 (a) and 4.6 (b) show the transient response of the MDSANF (Objective 1) Monte Carlo runs at the TT corner (T = 300.15 K, V_{DD} = 0.8 V) with ΔV_{TH} = 50 mV device-to-device variation. Each subplot corresponds to one of the four possible input combinations $AB \in \{00, 01, 10, 11\}$. In Fig. 4.6 (a), the MDSANF implements the AND/NAND function. The red band (SAout) discharges to 0 V for all cases except $AB = 11$, where it remains at the pre-charged level of 0.8 V; conversely, the blue band (SAoutb) shows the complementary NAND output. The thickness of each colored band indicates the spread of the Monte Carlo samples due to threshold-voltage mismatch. In Fig. 4.6 (b), the MDSANF implements the OR/NOR function. Here the red waveform only discharges for $AB = 00$, remaining high otherwise, while the blue waveform



(a) **(b)**

Figure 4.5: Transient response for compute mode of proposed 8TD SRAM cell for all possible input operand combinations (AB), '00,01,10,11' using (a) concurrent word-line and (b) sequential word-line activations at V_{DD} = 0.8 V.

produces the NOR complement. Note that the NOR band for $AB = 00$ is somewhat wider, reflecting slightly larger variation sensitivity in the all-zero case. Despite the 50 mV threshold perturbation, the sense amplifier cleanly resolves both true and complement outputs in under 0.5 ns for every input vector, demonstrating robust high-speed sensing and strong variation tolerance in large-scale Monte Carlo trials.

Figure 4.7 presents the derived statistical latency profiles obtained from the above Monte-Carlo simulations ($\Delta V_{TH} = \pm 50$ mV, $T = 300.15$ K, $V_{DD} = 0.8$ V) for every input combination of the operands A and B. The upper row depicts NOR operation, whereas the lower row shows NAND. For each case—'00', '01/10' and '11'—the measured latencies are binned into histograms; a Gaussian probability-density function fitted to the same data is super-imposed, illustrating that the delay distribution is well-approximated by a normal law driven primarily by random threshold-voltage mismatch. The coefficient of variation (CV $= \sigma/\mu$) rises from 0.7645 % for $AB = 00$ to 1.1570 % for the asymmetric inputs $AB = 01/10$, and peaks at 2.4873 % for $AB = 11$. This monotonic increase is consistent with the greater number of stacked pull-down devices conducting when both inputs are high, which amplifies the impact of mismatch on the internal discharge node. The mean latency likewise grows from $\mu = 381.4$ ps (00) to $\mu = 402.2$ ps (11), identifying the latter as the worst-case vector for NOR.

Figure 4.8 illustrates the Monte Carlo transient response of the proposed 8TD SRAM-IMC for XNOR and XOR realization under threshold-voltage variation. The results clearly follow the expected Boolean behavior: for equal input combinations, i.e., $AB = 00$ and $AB = 11$, the

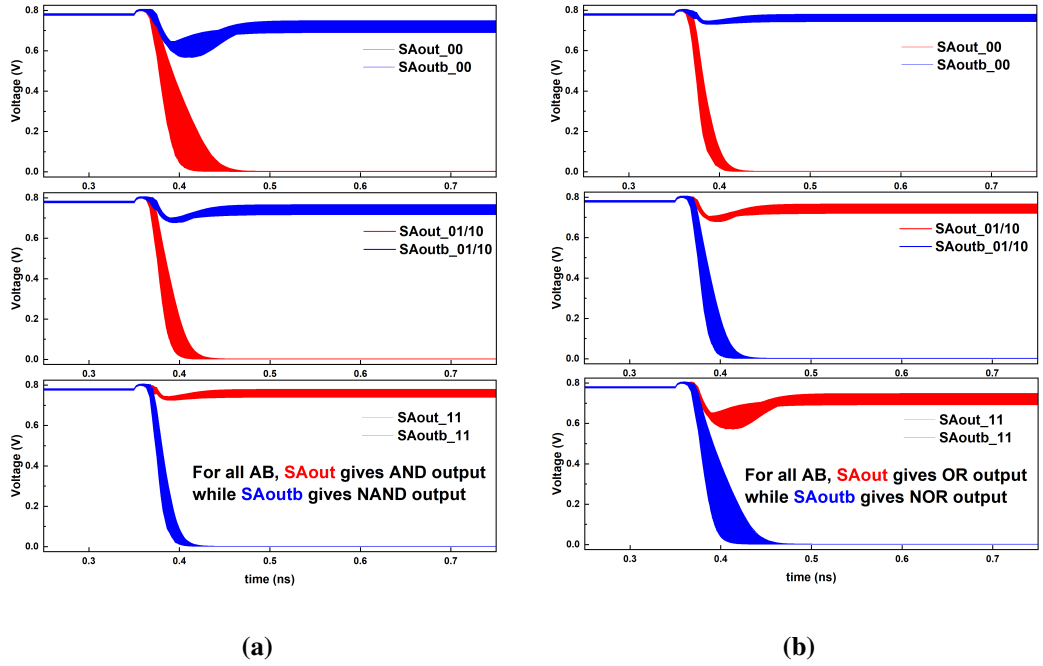


Figure 4.6: Transient response of the asymmetric sense amplifier for 6k Monte carlo simulations at TT corner for T = 300.15 K and V_{DD} = 0.8 V to implement (a) NAND and (b) NOR outputs for all possible input cases of input operands (AB), '00,01,10,11', at for ΔV_{TH} = 50 mV.

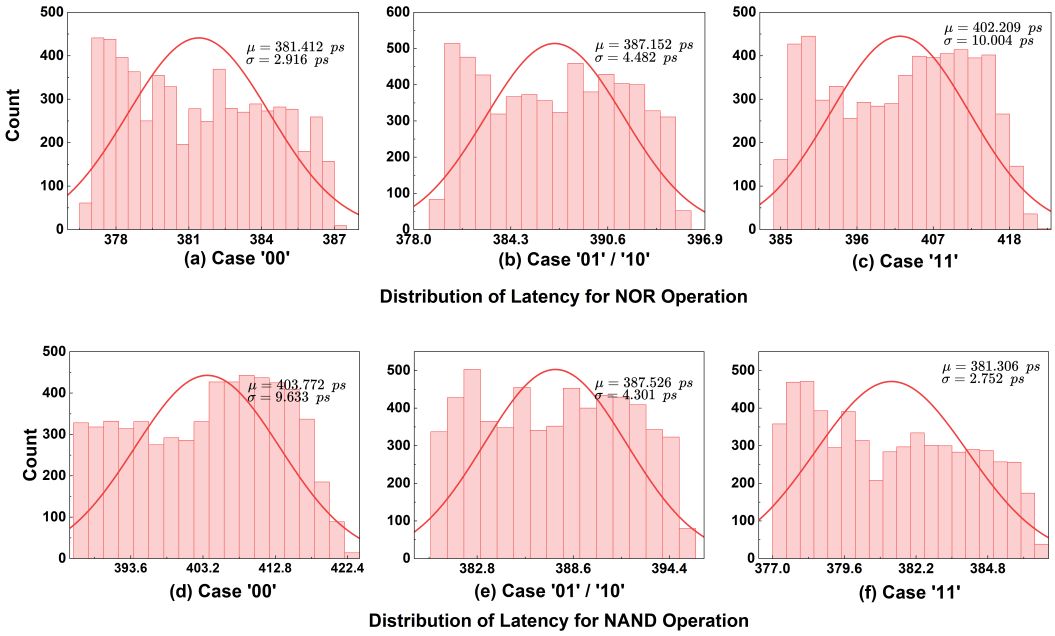


Figure 4.7: Distribution of latency for NOR and NAND logic operation for all possible combinations of input operands (AB), '00,01,10,11' for 6k Monte Carlo simulation runs at T = 300.15 K and V_{DD} = 0.8 V with ΔV_{TH} = 50 mV.

XNOR output remains at logic high while the XOR output stays near logic low, whereas for unequal input combinations, $AB = 01$ and $AB = 10$, the XNOR output discharges to logic low

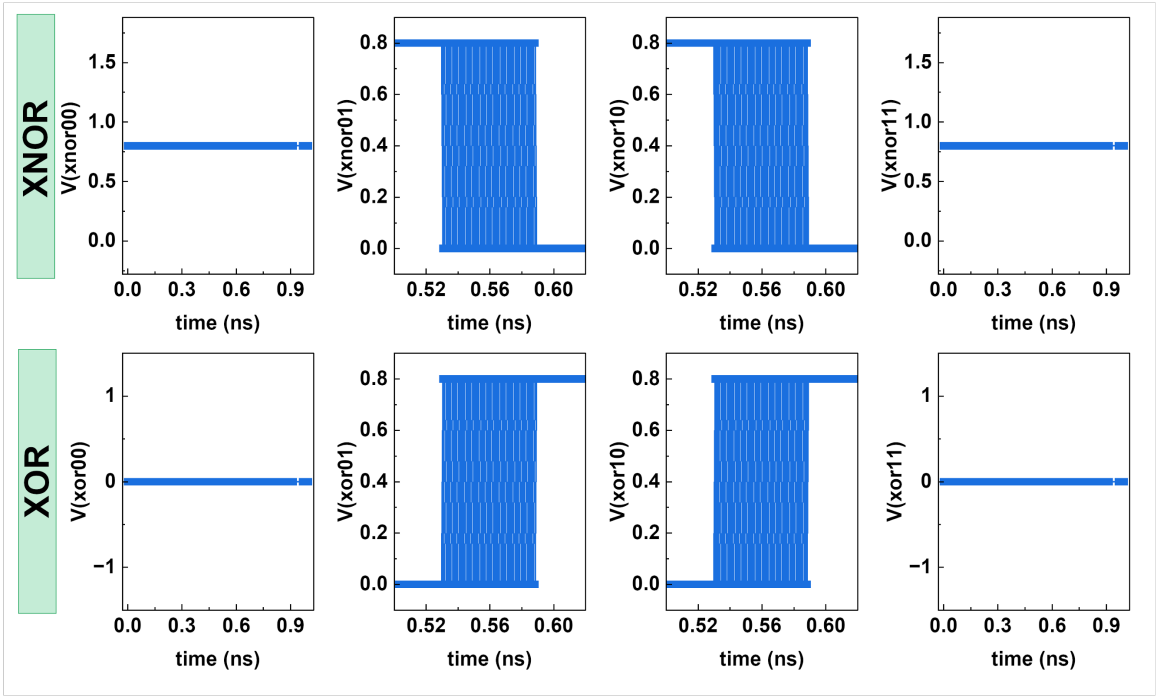
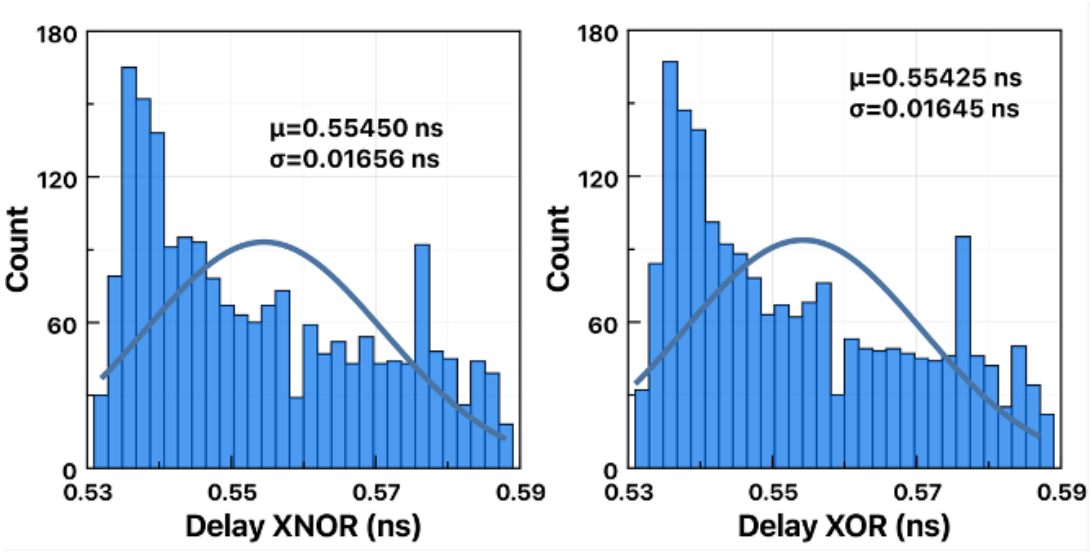


Figure 4.8: Transient simulations for (a) XNOR and (b) XOR logic operations for all possible input operand combinations, $AB = 00, 01, 10$, and 11. The results are obtained from 2000 Monte Carlo simulation runs at $T = 300.15$ K, $V_{\mathrm{DD}} = 0.8$ V, and $\Delta V_{\mathrm{TH}} = 50$ mV.

and the XOR output switches to logic high. The two mixed-input cases exhibit nearly identical transition bands, indicating good symmetry of the sensing mechanism, while the equal-input cases remain tightly clustered with negligible spread because no output transition is required. Overall, the separation between the logic-high and logic-low states across the Monte Carlo runs confirms that the proposed 8TD SRAM-IMC macro can reliably implement both XNOR and XOR functions at the TT corner with $\Delta V_{TH} = 50\,\mathrm{mV}$.

The trend reverses for NAND: the highest CoV of 2.3857 % occurs when $AB = 00$, because both pull-up devices must overcome the regenerative latch, while the lowest CoV of 0.7217 % is observed for $AB = 11$. Intermediate inputs $AB = 01/10$ yield a CoV of 1.1098 %. The largest mean latency appears again at the vector where both series-devices switch, i.e. $\mu = 403.4$ ps for $AB = 00$. Even in the worst cases the CoV stays below 2.5 %, and the spread ($\sigma \leq 10$ ps) remains a small fraction of the nominal ~ 0.4 ns delay, confirming that the asymmetric sense-amplifier maintains tight timing dispersion under significant threshold variation. The observation that NOR's slowest vector is $AB = 11$ while NAND's is $AB = 00$ provides a clear guideline for sizing: the pull-down stack in NOR and the pull-up stack in NAND should be marginally upsized if additional guard-band is required for six-sigma yield. Overall, the fitted Gaussian envelopes validate the robustness of the sensing strategy across all logic flavours and input patterns. Figure 4.9 presents the Monte Carlo latency distributions of the proposed 8TD SRAM-IMC macro for XNOR and XOR logic operations under device-threshold mismatch. The simulations include all four possible input combinations, $AB = 00, 01, 10$, and 11, thereby capturing the variation sensitivity of the complete logic-evaluation condition rather than a single data case.



(a) **(b)**

Figure 4.9: Latency distributions for (a) XNOR and (b) XOR logic operations for all possible input operand combinations, $AB = 00, 01, 10$, and 11. The results are obtained from 2000 Monte Carlo simulation runs at $T = 300.15$ K, $V_{DD} = 0.8$ V, and $\Delta V_{TH} = 50$ mV.

For the XNOR operation, the mean delay is 0.55450 ns with a standard deviation of 0.01656 ns, while the XOR operation exhibits a very close mean delay of 0.55425 ns with a standard deviation of 0.01645 ns. The difference between the two mean values is only about 0.25 ps, indicating that the complementary XOR/XNOR evaluation paths are well balanced.

The two distributions are concentrated mainly within the range of approximately 0.53–0.59 ns, showing limited delay dispersion even under $\Delta V_{TH} = 50$ mV. The standard deviation is nearly 3% of the mean delay for both operations, which suggests that the proposed differential SRAM-IMC structure maintains stable timing behavior in the presence of process-induced device mismatch. The close agreement between the XOR and XNOR distributions is expected because both functions are derived from the same differential compute mechanism, with the final logic interpretation determined by complementary bit-line responses. Therefore, the result confirms that the proposed macro can perform XOR/XNOR operations with almost identical latency characteristics and without introducing a significant timing penalty for either logic function.

4.5 PERFORMANCE COMPARISON OF IN-MEMORY SRAM CELL ARCHITECTURES

This section presents a comparative evaluation of the SRAM-IMC cell architectures discussed in the preceding sections. The comparison includes the conventional 6T SRAM-IMC structure, the conventional read-decoupled 8T cell (8TC), the differential 8T SRAM-IMC cell (8TB), the 8TE SRAM-IMC baseline, and the proposed 8TD SRAM-IMC cell. All cells are implemented using the Predictive Technology Model (PTM) at the 32 nm technology node [145]. To ensure a consistent comparison, identical word-line timing, sense-amplifier enable timing, device sizing methodology, and bit-line loading conditions are used across the evaluated architectures.

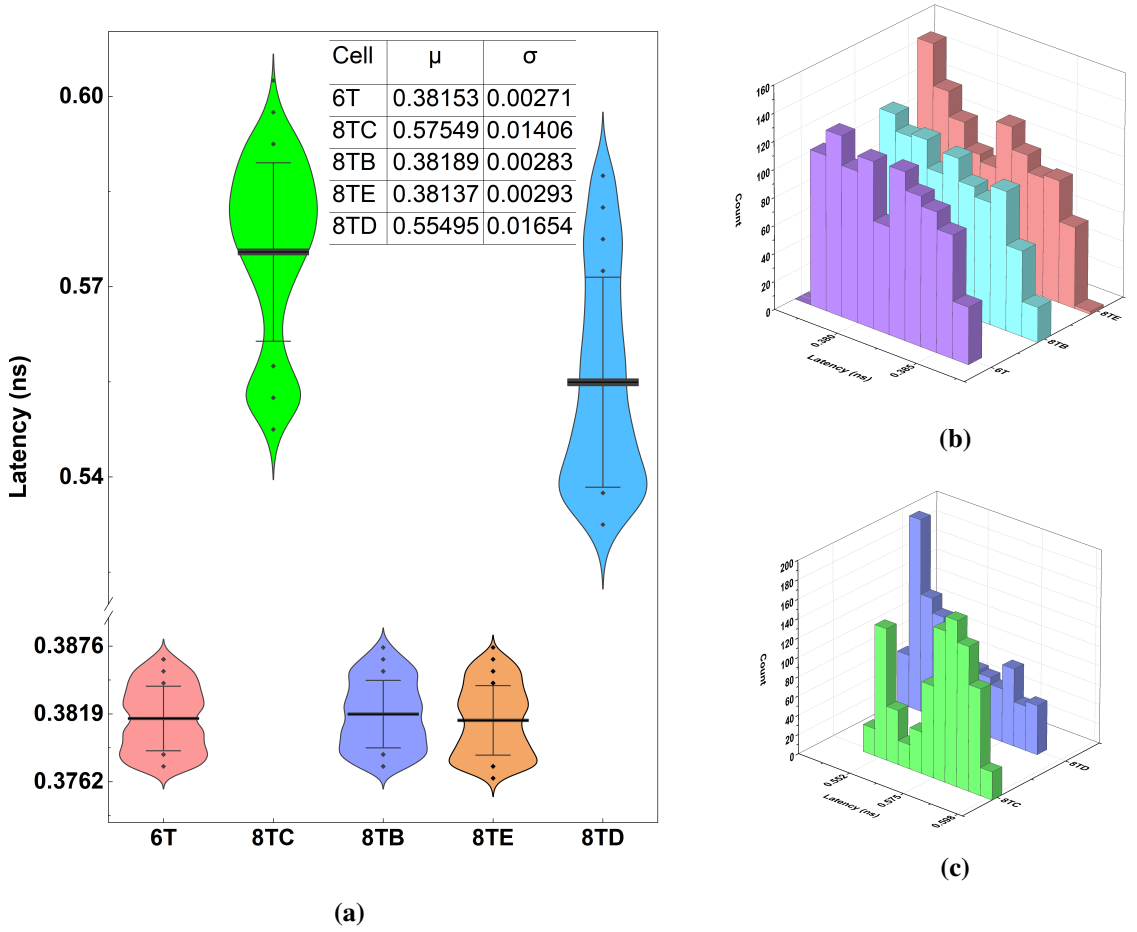


Figure 4.10: SRAM-IMC latency distributions under device mismatch at $V_{DD} = 0.8$ V and $\Delta V_{TH} = 50$ mV: (a) violin-plot summary with extracted μ and σ, (b) Monte Carlo latency distribution for the fast cluster consisting of 6T, 8TB, and 8TE cells, and (c) Monte Carlo latency distribution for the higher-latency 8TC and 8TD cells.

The supply voltage is swept from $V_{DD} = 0.5$ V to 1.3 V, covering both near-threshold and nominal-voltage operation. The bit-line capacitance is modeled using a π-network representation to emulate large-array loading, with $C_{BL} = 50$ fF used for the comparative evaluation.

The primary comparison metrics are latency, latency spread under threshold-voltage variation, average energy per bit, throughput, and energy reduction with respect to reference SRAM-IMC macros. These metrics are selected because SRAM-IMC design involves a coupled trade-off between speed, energy, variation tolerance, and peripheral complexity. A cell that provides low latency but large energy overhead may not be suitable for energy-constrained in-memory computing. Conversely, a cell that achieves very low energy but suffers large delay may not meet throughput requirements. Therefore, the objective of this comparison is to identify the practical operating position of each cell architecture across supply voltage and process variation.

Figure 4.10 compares the latency distributions of the evaluated SRAM-IMC cells under device mismatch at $V_{DD} = 0.8$ V and $\Delta V_{TH} = 50$ mV. The violin plot in Fig. 4.10(a) provides a compact view of both the mean latency and the statistical spread, while Figs. 4.10(b)–(c) show

the corresponding 3D Monte Carlo distributions.

The 6T, 8TB, and 8TE cells form the fastest latency group. Their mean latencies are 0.38153 ns, 0.38189 ns, and 0.38137 ns, respectively. The difference among these three cells is below 0.6 ps, indicating that all three architectures operate within almost the same timing band under the chosen sensing conditions. Their standard deviations are also small: 0.0027 ns for 6T, 0.0028 ns for 8TB, and 0.0029 ns for 8TE. This corresponds to coefficients of variation of approximately 0.71%, 0.73%, and 0.76%, respectively, confirming that the fast group also maintains tight delay dispersion under mismatch.

The 8TC cell exhibits the largest mean latency, 0.57549 ns, and a much larger standard deviation of 0.0140 ns. Compared with the 8TE cell, the 8TC latency is approximately 50.9% higher. This is consistent with its single-ended read/compute path, where the sensing operation depends on the development of a single read bit-line level relative to a reference or switching threshold. Although this architecture is attractive from an energy perspective, its delay and delay spread are significantly larger than the differential-sensing-based structures.

The proposed 8TD cell reduces the mean latency relative to 8TC from 0.57549 ns to 0.55495 ns, corresponding to an improvement of approximately 3.6%. However, its standard deviation remains relatively large at 0.0165 ns, giving a coefficient of variation of approximately 3.0%. Thus, the 8TD cell improves the practical latency position relative to 8TC but still exhibits greater timing spread than the fast differential cluster. This indicates that the proposed 8TD architecture is better interpreted as an energy-oriented improvement over 8TC rather than as a pure latency replacement for the 6T/8TB/8TE group.

The 3D Monte Carlo distributions further support this interpretation. The distributions in Fig. 4.10(b) are concentrated near the 0.38 ns region for 6T, 8TB, and 8TE, whereas Fig. 4.10(c) shows wider and higher-latency distributions for 8TC and 8TD. The separation between the two clusters demonstrates that the sensing strategy and read-path organization play a dominant role in determining compute latency under mismatch. Differential sensing provides strong timing control, while single-ended or modified single-ended evaluation paths generally reduce energy at the cost of larger delay and delay spread.

Figure 4.11 compares the latency, average energy per bit, and throughput of the SRAM-IMC cells across the supply-voltage range. The latency trend in Fig. 4.11(a) shows that all cells become faster as V_{DD} increases. This is expected because higher supply voltage increases device overdrive, strengthens the bit-line discharge or charge path, and reduces the time required for the sensing circuitry to resolve the compute output. The largest latency reduction occurs between 0.5 V and 0.8 V, where the circuits move away from near-threshold operation. Beyond approximately 0.9 V, the latency curves flatten, indicating diminishing speed improvement with further supply increase.

The 8TC cell remains the slowest architecture across the complete supply-voltage range. At V_{DD} = 0.8 V, its latency is approximately 0.575 ns, whereas the 6T, 8TB, and 8TE cells operate close to 0.381 ns. Therefore, the 8TC delay penalty is not limited to a single supply voltage but persists throughout the sweep. This behavior is a consequence of the single-ended read path and

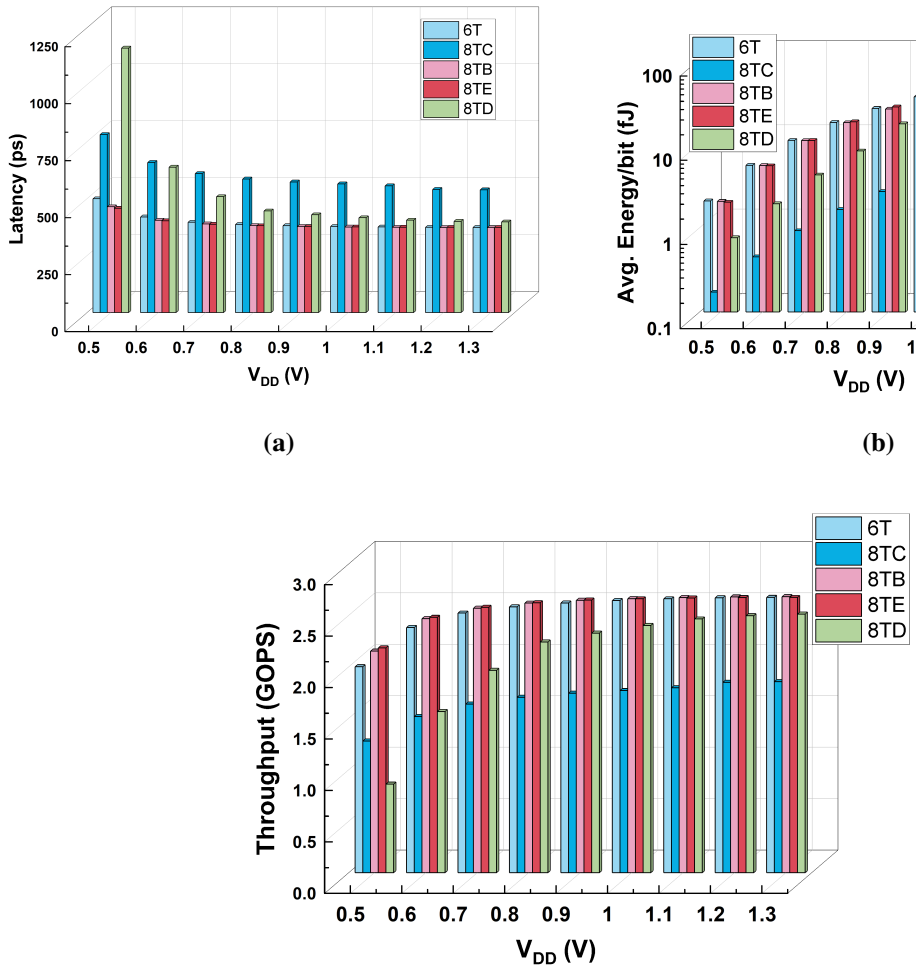


Figure 4.11: Performance comparison of SRAM-IMC cell architectures as a function of V_{DD} at the TT process corner and $T = 300.15$ K: (a) latency, (b) average energy per bit, and (c) throughput for 6T, 8TC, 8TB, 8TE, and 8TD SRAM-IMC cells.

its dependence on slower voltage development at the read bit-line. In contrast, the 6T, 8TB, and 8TE structures remain grouped together because their compute operation is resolved through stronger bit-line discharge and differential or differential-like sensing. The 8TD cell lies between the 8TC and fast-cluster designs: it improves over 8TC, particularly at lower and mid supply voltages, but it does not fully reach the latency of the fastest differential group.

The average energy per bit in Fig. 4.11(b) follows the opposite trend. Energy increases with V_{DD} for all architectures, which is consistent with the dynamic switching-energy dependence $E \propto CV_{DD}^2$. The 8TC cell consumes the least energy per bit across the supply range, reflecting the lower switched capacitance and reduced single-ended activity in its read path. However, this energy advantage is obtained at the cost of the highest latency and lowest throughput. The 6T, 8TB, and 8TE cells consume higher energy because their compute evaluation involves stronger bit-line activity and differential sensing. The proposed 8TD cell achieves a more favorable

energy position than the fast differential group while remaining faster than 8TC, thereby shifting the practical speed–energy trade-off toward a more balanced operating point.

The throughput results in Fig. 4.11(c) follow directly from the latency trends. Since throughput is inversely related to compute latency, the 6T, 8TB, and 8TE cells provide the highest throughput across most of the voltage range. At V_{DD} = 0.8 V, the 8TE cell achieves approximately 2.624 GOPS, while 8TB and 6T provide approximately 2.620 GOPS and 2.584 GOPS, respectively. The 8TC cell provides only about 1.706 GOPS at the same supply voltage because of its higher delay. The 8TD cell improves the throughput relative to 8TC and approaches the faster group more closely as the supply voltage increases. This again confirms that the proposed 8TD cell should be understood as a compromise architecture that improves the energy efficiency of the SRAM-IMC operation while recovering part of the speed lost by 8TC.

The combined interpretation of Fig. 4.11 is that no single cell dominates all metrics. The 8TC cell is energy efficient but slow. The 6T cell is fast and compact, but simultaneous word-line activation can produce read disturb and reliability concerns in IMC operation. The 8TB cell improves reliability through differential operation but introduces additional transistor and control overhead. The 8TE cell provides the best latency and throughput among the evaluated fast architectures while maintaining modest area overhead. The proposed 8TD cell offers an improved energy position relative to the fast differential baselines while achieving lower latency than 8TC. Thus, the choice of cell depends on whether the design target prioritizes maximum speed, minimum energy, variation tolerance, or a balanced speed–energy operating point.

Figure 4.12 summarizes the energy advantage of the proposed 8TD SRAM-IMC cell with respect to the 6T, 8TB, and 8TE SRAM-IMC baselines. The heatmap shows that the proposed 8TD cell provides the largest energy reduction in the low-to-mid supply-voltage region. At V_{DD} = 0.5–0.7 V, the reduction is typically in the range of 60%–70%, indicating that the proposed cell is particularly effective under energy-constrained near-threshold operation. At V_{DD} = 0.8 V, the reduction remains strong, with values near the mid-50% range for most reference macros. In the nominal region of approximately 0.9–1.1 V, the energy benefit decreases but remains significant, typically around 30%–40%. At higher supply voltages, 1.2–1.3 V, the reduction tapers to approximately 15%–25%.

The reduction decreases at higher V_{DD} because all cells experience a stronger dynamic-energy increase as the supply voltage rises. As the V_{DD}^2 component becomes dominant, the relative advantage of reducing capacitance or modifying the compute path becomes smaller. Nevertheless, the 8TD cell maintains a positive energy benefit across the entire voltage range. This is important because it confirms that the proposed architecture does not provide an isolated advantage at one operating point; rather, it consistently reduces compute energy over a broad supply range.

The heatmap also shows that the energy advantage is comparable across the 6T, 8TB, and 8TE baselines, although the exact reduction varies slightly with the reference architecture and supply voltage. This suggests that the proposed 8TD cell reduces the effective switching activity of the compute operation in a way that is not limited to one particular comparison. In other words, the energy improvement is not simply the result of selecting a weak reference cell; it remains visible

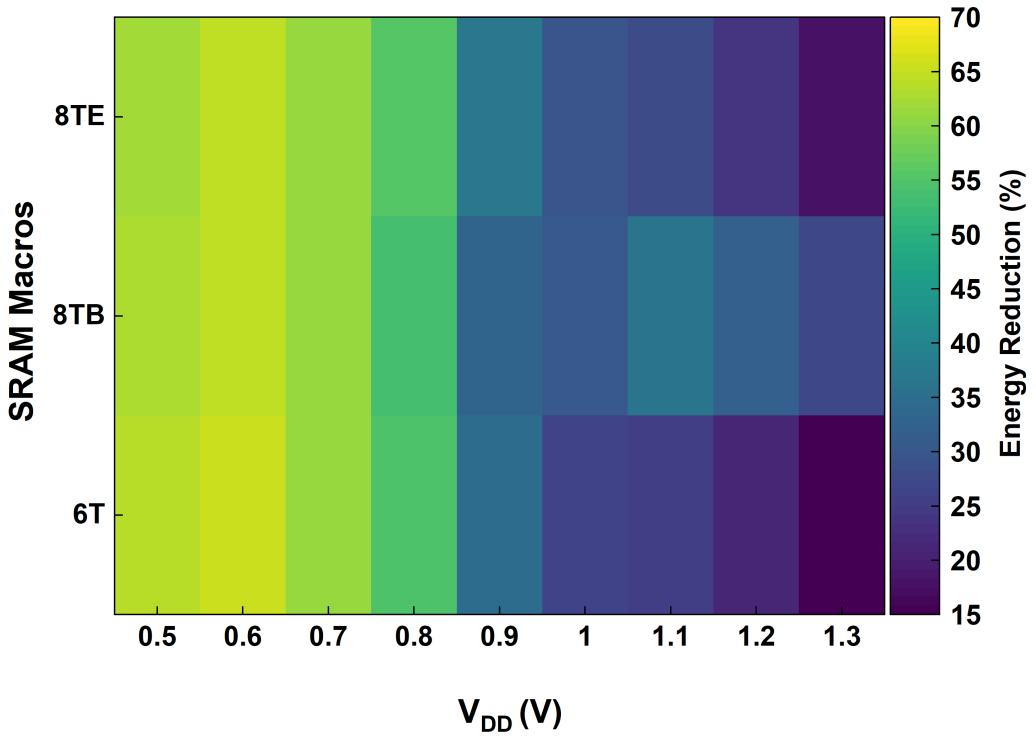


Figure 4.12: Energy reduction of the proposed 8TD SRAM-IMC cell with respect to 6T, 8TB, and 8TE SRAM-IMC macros across V_{DD}.

even when compared against the faster and more competitive 8TB and 8TE structures.

Table 4.1 summarizes the latency statistics extracted from the mismatch study in Fig. 4.10. The values reinforce the clustering visible in the violin plot. The 6T, 8TB, and 8TE cells exhibit nearly identical mean latency and low statistical spread. The 8TC and 8TD cells have higher mean latency and larger spread, but the proposed 8TD still reduces the mean delay relative to 8TC. Therefore, the latency statistics confirm that the proposed 8TD architecture improves the speed of the energy-oriented 8TC-style operating regime, while the 8TE cell remains the best option when maximum latency and throughput performance are prioritized.

Table 4.2 provides a representative operating-point comparison at V_{DD} = 0.8 V. The table highlights the different roles of the evaluated cells. The 8TC cell achieves the lowest energy per bit, 1.647 fJ, but it also has the highest latency, 586.166 ps, and the lowest throughput, 1.706 GOPS. The 8TB and 8TE cells provide the lowest latency and highest throughput, with the 8TE cell reaching 381.097 ps and 2.624 GOPS. Compared with 8TC, the 8TE cell reduces latency by approximately 35.0% and improves throughput by approximately 53.8%, while maintaining much lower normalized area than 8TB. The 6T cell is also competitive in speed and area, but its IMC reliability is constrained by the shared read/write path and the possibility of read disturb under multi-row activation.

The overall performance comparison leads to three important conclusions. First, the 6T, 8TB, and 8TE cells form the speed-oriented group, with low latency, high throughput, and

Table 4.1: Latency statistics of SRAM-IMC cells under device mismatch at V_{DD} = 0.8 V and ΔV_{TH} = 50 mV.

Cell	μ (ns)	σ (ns)	CV (%)
6T	0.38153	0.0027	0.71
8TC	0.57549	0.0140	2.43
8TB	0.38189	0.0028	0.73
8TE	0.38137	0.0029	0.76
8TD	0.55495	0.0165	2.97

narrow delay spread. Second, the 8TC cell forms the energy-oriented baseline, achieving the lowest average energy per bit but suffering a large speed penalty. Third, the proposed 8TD cell improves the energy profile relative to the fast reference macros and reduces latency relative to 8TC, making it useful for applications where energy efficiency is more important than absolute maximum speed.

The layout comparison in Fig. 4.13 provides the physical basis for the normalized area values reported in Table 4.2. The 6T cell has the smallest normalized area and is therefore the density baseline. The 8TC and 8TB cells introduce additional transistors and routing complexity due to their read-decoupled and differential structures, resulting in higher normalized area. The 8TE cell provides a more favorable physical trade-off: it retains the latency and throughput advantage of the fast differential group while requiring only a modest area overhead over the 6T baseline. This makes it attractive for speed-oriented SRAM-IMC operation where the added logic capability must not significantly compromise array density.

Overall, the comparison establishes two complementary SRAM-IMC design directions. The 8TE cell provides the strongest latency-throughput position among the evaluated reference cells while maintaining only modest area overhead. The 8TC cell provides the lowest energy per bit but suffers a clear delay and throughput penalty. The proposed 8TD cell improves the energy profile relative to the faster 6T, 8TB, and 8TE baselines, while reducing the latency penalty

Table 4.2: Representative SRAM-IMC performance at V_{DD} = 0.8 V.

SRAM-IMC	Latency (ps)	Avg. Energy/bit (fJ)	Throughput (GOPS)	Normalized Area
6T	386.982	17.805	2.584	1.000
8TC	586.166	1.647	1.706	1.264
8TB	381.675	17.628	2.620	1.394
8TE	381.097	18.0895	2.624	1.065
8TD	445.920	8.136	2.242	1.067

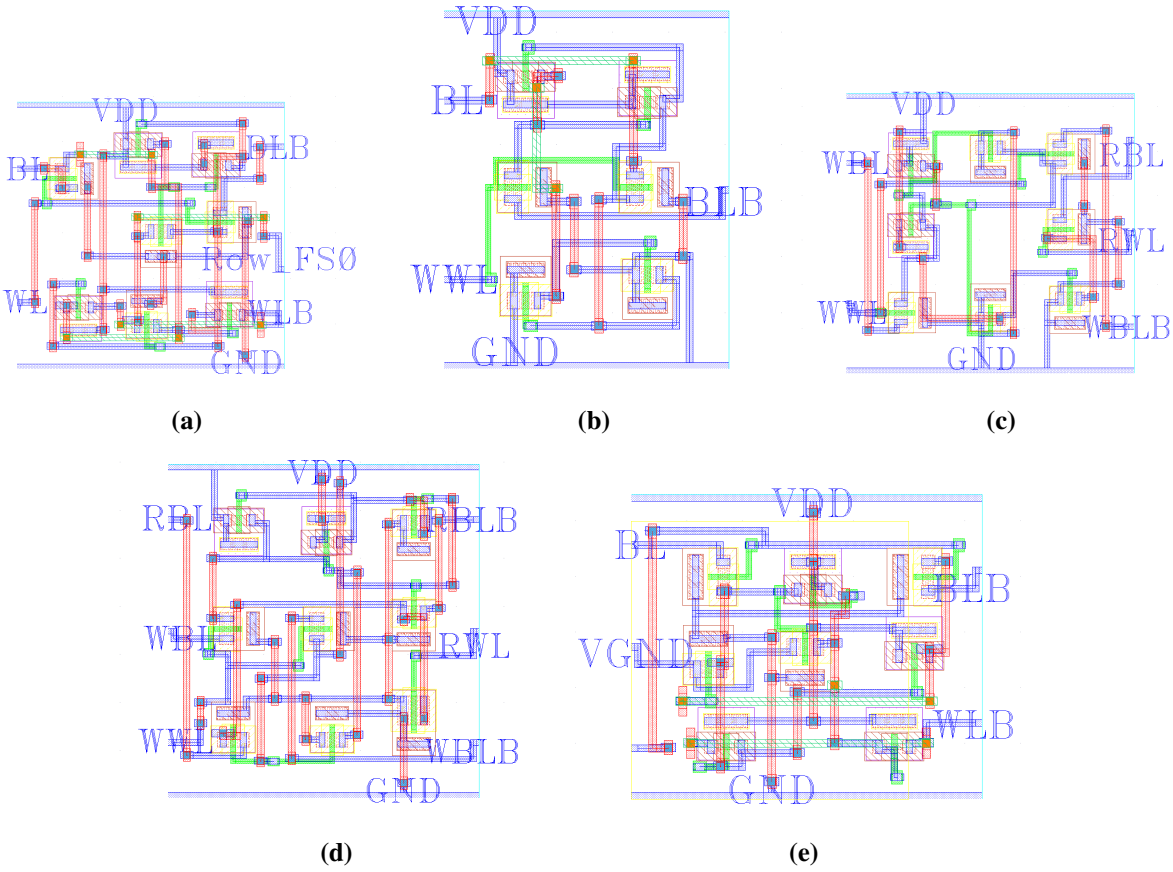


Figure 4.13: Layout views of the SRAM cells considered for IMC implementation: (a) 8TE, (b) 6T, (c) 8TC, (d) 8TB and (e) 8TD. The layouts are used to support the normalized area comparison reported in Table 4.2.

associated with 8TC. Therefore, the 8TE structure is better suited for speed-oriented IMC operation, whereas the proposed 8TD cell is more suitable when energy reduction is prioritized without accepting the full speed penalty of conventional 8TC-based IMC.

4.6 SUMMARY OF IMPORTANT RESULTS

This chapter presented SRAM-based in-memory computation for bitwise logic implementation using conventional and proposed SRAM cell structures. The proposed 8TD SRAM-IMC macro was evaluated for NAND/NOR and XOR/XNOR operation, and the resulting designs were compared in terms of latency, energy, throughput, variation tolerance, and area. The important results are summarized below.

1. Conventional 6T, 8TC, 8TB, and 8TE SRAM-IMC structures show distinct speed–energy–reliability trade-offs. The 6T cell is compact and fast but may suffer from read disturb under multi-row activation, whereas 8TC improves read isolation and energy efficiency at the cost of higher latency. The 8TB and 8TE cells provide stronger differential operation and better timing behavior.
2. The proposed 8TD SRAM-IMC macro realizes NAND/NOR and XOR/XNOR logic by

connecting two SRAM cells in a column-wise configuration. The stored operands A and B interact through the shared bit-line and sensing path, allowing Boolean computation to be performed inside the memory structure without transferring data to external logic.

3. Functional simulations confirm correct logic operation for all input combinations, $AB = 00, 01, 10$, and 11. Both concurrent and sequential word-line activation schemes were analyzed, showing that the proposed structure supports logic evaluation while controlled activation helps reduce read-disturb concerns.
4. Monte Carlo analysis confirms robust NAND/NOR operation under mismatch. For 6000 runs at $V_{DD} = 0.8$ V, $T = 300.15$ K, and $\Delta V_{TH} = 50$ mV, the outputs remain clearly separated for all input cases. The worst-case mean delay remains close to 0.4 ns, and the coefficient of variation stays below approximately 2.5%.
5. XOR/XNOR operation also remains stable under mismatch. For 2000 Monte Carlo runs, the XNOR and XOR mean delays are 0.55450 ns and 0.55425 ns, respectively, with standard deviations close to 0.0165 ns. The very small mean-delay difference confirms that the complementary XOR/XNOR paths are well balanced.
6. The mismatch-based latency comparison separates the cells into two timing groups. The 6T, 8TB, and 8TE cells form the fast cluster with mean delays near 0.381 ns, while 8TC and 8TD form the higher-latency group. The proposed 8TD still improves over 8TC by reducing the mean delay from 0.57549 ns to 0.55495 ns.
7. The voltage-sweep comparison shows that increasing V_{DD} reduces latency and improves throughput, while average energy per bit increases. At $V_{DD} = 0.8$ V, the 8TE cell gives the best speed-oriented result with 381.097 ps latency and 2.624 GOPS throughput, whereas 8TC gives the lowest energy per bit but the highest latency.
8. The proposed 8TD cell provides consistent energy reduction with respect to 6T, 8TB, and 8TE baselines. The reduction reaches approximately 60%–70% at $V_{DD} = 0.5$–0.7 V, remains around 30%–40% near nominal supply, and tapers to approximately 15%–25% at higher supply voltages.
9. The layout and normalized-area comparison show that 6T remains the density baseline, while 8TC and 8TB introduce larger area overhead due to additional read and differential paths. The 8TE cell achieves a favorable physical trade-off by retaining high latency-throughput performance with only modest area overhead.

In summary, the chapter shows that SRAM-IMC performance depends on the combined choice of bit-cell structure, sensing strategy, and supply voltage. The proposed 8TD SRAM-IMC macro provides robust NAND/NOR and XOR/XNOR computation under mismatch, while the comparative analysis identifies 8TE as preferable for speed-oriented operation and 8TD as preferable for energy-oriented operation.

CHAPTER – 5

RECONFIGURABLE SRAM-IMC ARRAY FOR THROUGHPUT ENHANCEMENT

5.1 OVERVIEW OF THE CHAPTER

The preceding chapters established the three circuit-level foundations required for an SRAM-based in-memory computing (SRAM-IMC) platform: a modified differential sensing strategy, a robust 8T SRAM bit-cell, and an SRAM-IMC macro capable of Boolean logic evaluation. While the cell-level and macro-level results demonstrate improvements in stability, latency, energy, and variation tolerance, a practical SRAM-IMC system must also preserve these advantages when the design is extended to an array-level organization. Array scaling introduces longer bit-lines, larger parasitic capacitance, higher interconnect loading, and stronger interaction between the selected cells and peripheral circuitry. Consequently, the design challenge moves beyond demonstrating correct logic functionality in a small macro; the complete array must support reliable memory access and compute-mode evaluation under the same physical bit-line and sensing constraints.

Recent SRAM-CIM and SRAM-IMC designs have shown that configurable SRAM macros can support logic, arithmetic, and machine-learning-oriented computation by reusing the memory array for bit-line processing, precision-configurable operation, and column-parallel data paths [84, 88, 104, 106]. However, array-level reconfigurability also introduces new trade-offs. The same peripheral circuits must satisfy the requirements of conventional memory operation, where static noise margin, write margin, read delay, and read power are critical, as well as compute operation, where latency, average energy per bit, and throughput determine the usefulness of the IMC mode. Therefore, the design of a reconfigurable SRAM-IMC array requires co-optimization of the bit-cell, sense amplifier, timing-control scheme, and compute data path rather than independent optimization of each block [7].

In this chapter, the modified sense-amplifier strategy, the proposed 8T SRAM cell, and the SRAM-IMC logic macro developed in the preceding chapters are integrated into a unified dual-mode architecture. The proposed architecture operates as a conventional SRAM array in memory mode and as a logic-enabled SRAM-IMC array in compute mode. In memory mode, the selected cell develops a differential bit-line signal that is resolved by the sense amplifier to generate full-swing read outputs. In compute mode, the same bit-line and sensing infrastructure is reused to evaluate Boolean logic functions through the IMC path. This shared organization reduces peripheral duplication and establishes a direct path from circuit-level improvement to array-level throughput enhancement.

In light of the preceding requirement, the objective of this chapter is articulated as follows:
"Design and optimization of a reconfigurable SRAM-IMC array using modified sense amplifier and computation macro for throughput enhancement"

Methodology

1. Develop a dual-mode SRAM-IMC array architecture that integrates the modified sensing strategy, the proposed 8T SRAM cell, and the Boolean compute macro into a shared memory/compute data path.

2. Define the required control sequence for memory-mode and compute-mode operation, including precharge, word-line activation, complementary word-line control, virtual-ground control, and sense-amplifier enable timing.

3. Evaluate memory-mode scalability for array sizes from 2×2 to 32×32 across a supply-voltage range of $V_{DD} = 0.6\,\text{V}$ to $1.2\,\text{V}$, using static noise margin, write-margin deviation, read delay, and read power as the key metrics.

4. Evaluate compute-mode scalability using NAND-evaluation latency, average energy per bit, and throughput across the same array-size and supply-voltage range.

5. Identify the dominant scaling penalties in memory and compute modes, and use these results to assess the suitability of the proposed reconfigurable SRAM-IMC array for throughput-enhanced operation.

The chapter is organized as follows. Section 5.2 introduces the need for reconfigurable SRAM-IMC arrays and motivates the transition from cell-level and macro-level design to array-level evaluation. Section 5.3 presents the proposed dual-mode SRAM-IMC architecture and explains the shared data path used for memory and compute operation. Section 5.4 analyzes the memory-mode behavior of the proposed array in terms of SNM, write-margin deviation, read delay, and read power. Section 5.5 evaluates the compute-mode latency, average energy per bit, and throughput under array-size and supply-voltage scaling. Finally, Section 5.6 summarizes the key findings and highlights the dominant trade-offs in the proposed reconfigurable SRAM-IMC array.

5.2 INTRODUCTION

The SRAM-IMC macro developed in Chapter 4 demonstrated that Boolean logic can be evaluated directly inside the SRAM data path using the proposed differential 8T SRAM cell and modified sensing strategy. At the macro level, the proposed structure was shown to support NAND/NOR and XOR/XNOR operation for all possible input combinations of the stored operands. The corresponding Monte Carlo analysis further confirmed that the compute outputs remain distinguishable under threshold-voltage mismatch. However, a macro-level demonstration does not fully establish the suitability of the design for practical memory-compute integration. Once the same compute mechanism is embedded within an array, the selected cells interact with longer bit-lines, larger parasitic capacitance, precharge devices, row-selection circuitry, timing-control signals, and shared sense-amplifier paths. Therefore, the design must be evaluated beyond the isolated cell or two-cell macro level to determine whether the proposed SRAM-IMC approach remains reliable when extended to an array-level organization.

In this chapter, the term reconfigurable SRAM-IMC array refers to an SRAM architecture that can operate in two modes using a largely shared circuit infrastructure. In memory mode, the array behaves as a conventional SRAM: the row decoder selects the required word line, the selected cell develops a differential bit-line signal, and the sense amplifier resolves the stored data. In compute mode, the same stored operands, bit-line pair, and sensing path are reused to evaluate Boolean logic locally inside the memory array. Thus, reconfigurability in this work does not imply a separate programmable logic fabric added beside the SRAM. Instead, it refers to the ability of the same memory array and peripheral data path to support both storage and in-memory computation with minimal duplication of circuitry.

This dual-mode requirement is consistent with the broader development of SRAM-based compute-in-memory and in-memory computing architectures. Early configurable SRAM designs showed that conventional SRAM arrays could be reused for search and logic operations through suitable word-line control and sensing techniques [5]. Logic-centric SRAM-IMC designs such as X-SRAM further demonstrated that Boolean operations can be performed inside CMOS SRAM by modifying the access scheme and sensing strategy [6]. Subsequent SRAM-CIM macros extended this idea toward vector computing, configurable precision, and neural-network-oriented workloads using bit-serial arithmetic, configurable unit macros, and multi-bit input/output support [84, 88, 106]. These works establish that SRAM arrays can be adapted for computation, but they also show that the cell, bit-line, sense amplifier, and control circuitry must be treated as a coupled system rather than as independent design blocks.

Recent SRAM-CIM and SRAM-IMC works further emphasize the importance of multi-mode and reconfigurable array operation. A 9T-SRAM IMC macro has been reported for Boolean logic and multiply-and-accumulate operation, illustrating the trend toward combining logic and arithmetic functionality within the same memory-oriented structure [197]. A configurable IMC architecture based on dual-port SRAM has also been proposed to exploit the flexibility of multi-port access and bit-line accumulation for compute operation [198]. Other recent designs have explored 9T SRAM-based CIM with capacitance weighting and redundant-array-assisted conversion, PVT-insensitive 7T SRAM-CIM using dynamic matching quantization, and 1–16b reconfigurable 7T SRAM-based digital near-memory computing for neural-network processing [199–201]. These developments indicate that SRAM-CIM research is moving beyond isolated Boolean demonstrations toward configurable array-level architectures that must balance memory reliability, compute speed, energy efficiency, and peripheral complexity.

Array-level evaluation is essential because the dominant design constraints change as the circuit is scaled from a single cell to a complete array. At the bit-cell level, static noise margin, write margin, read delay, and cell-level power are the primary reliability and performance metrics. At the macro level, the main concern is whether the bit-line response can be correctly sensed for each operand combination. At the array level, however, bit-line capacitance, word-line loading, interconnect parasitics, precharge energy, sense-amplifier offset, and peripheral timing begin to dominate the overall behavior. Prior SRAM array studies have shown that array organization, statistical variation, and bit-line loading directly influence yield, access time, and

energy efficiency [2, 8, 73, 74]. These challenges are even more critical in SRAM-IMC because the same array must satisfy both conventional memory-mode requirements and compute-mode logic-evaluation requirements [7].

Accordingly, this chapter integrates the modified sense-amplifier concept, the proposed differential 8T SRAM cell, and the SRAM-IMC logic macro into a unified reconfigurable array framework. The proposed architecture uses the same bit-line and sensing resources for both modes of operation. In memory mode, the array is evaluated in terms of static noise margin, write-margin deviation, read delay, and read power. In compute mode, the array is evaluated in terms of NAND-evaluation latency, average energy per bit, and throughput. The analysis is carried out for array sizes ranging from 2×2 to 32×32 over a supply-voltage range of V_{DD} = 0.6–1.2 V. This evaluation range allows the influence of array scaling to be studied together with voltage scaling, which is important for assessing the feasibility of the proposed architecture under both low-voltage and nominal-voltage operating conditions.

The contribution of this chapter is therefore twofold. First, it demonstrates how the circuit blocks developed in the preceding chapters can be combined into a dual-mode SRAM-IMC array that supports both memory access and in-memory Boolean computation through a shared data path. Second, it identifies the dominant scaling trends that appear when the proposed architecture is extended from small arrays to larger organizations. In particular, the analysis distinguishes between metrics that remain weakly affected by array size, such as static stability, and metrics that become more strongly dependent on array expansion, such as read power and compute energy. This distinction is important because an SRAM-IMC architecture is only useful if the compute capability is achieved without sacrificing the basic memory function. The following sections therefore present the proposed array architecture and then analyze its memory-mode and compute-mode behavior under voltage and array-size scaling.

5.3 PROPOSED RECONFIGURABLE SRAM-IMC ARRAY ARCHITECTURE

Figure 5.1 shows the proposed reconfigurable SRAM-IMC array architecture. The architecture combines the three major circuit-level blocks developed in the preceding chapters: the modified sense-amplifier topology, the proposed differential 8T SRAM cell, and the SRAM-IMC logic macro. The objective of this integration is to construct a dual-mode array that can operate either as a conventional SRAM for data storage and readout or as an in-memory computing array for Boolean logic evaluation. Instead of using separate memory and compute hardware paths, the proposed architecture reuses the same bit-line pair, row-control network, timing-control signals, and sensing circuitry for both modes of operation. This shared organization forms the basis of the reconfigurable SRAM-IMC framework analyzed in this chapter.

The proposed array consists of differential 8T SRAM cells arranged in a row-column structure. Each column is associated with a complementary bit-line pair, BL/BLB, and each row is accessed through word-line control generated by the row decoder. The row decoder selects the required row during memory access and also supports the controlled activation required during compute-mode operation. The timing-control block generates and coordinates the key signals

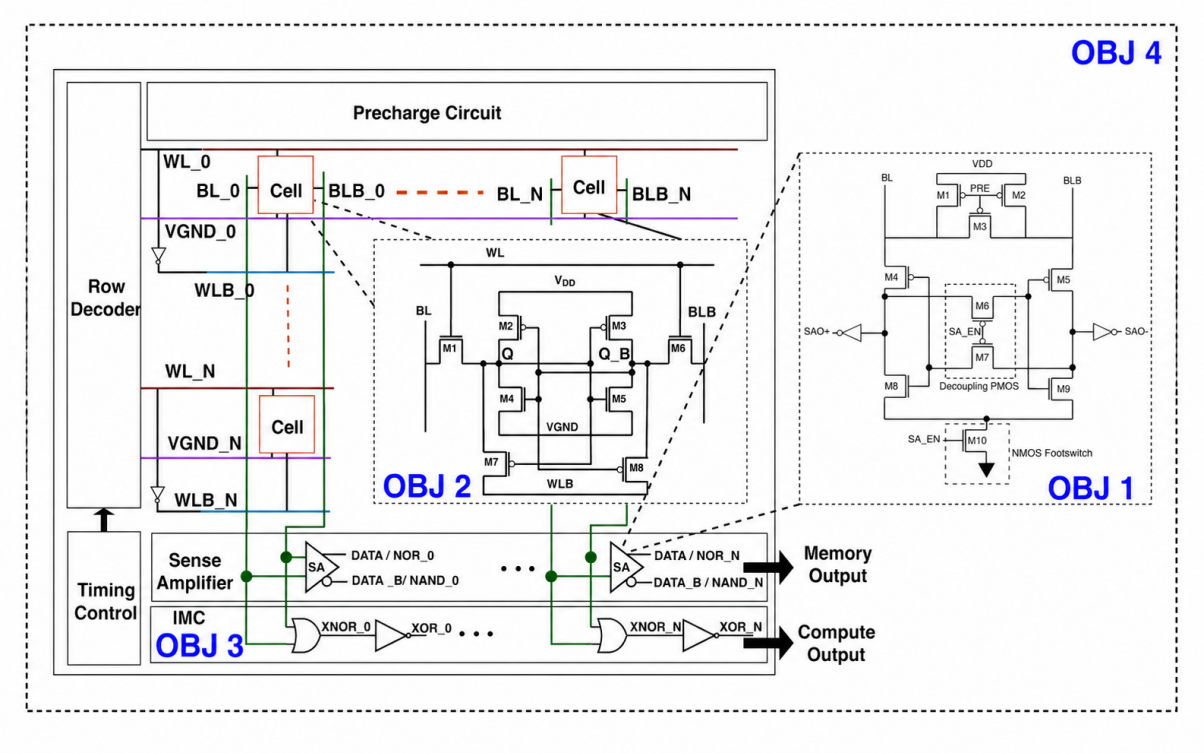


Figure 5.1: Proposed reconfigurable SRAM-IMC architecture showing the shared memory-mode and compute-mode data paths.

required by the proposed cell and sensing structure, including WL, WLB, VGND, precharge control, and sense-amplifier enable. Before each operation, the precharge circuit initializes BL and BLB to a defined voltage level, so that the subsequent bit-line development depends only on the selected storage nodes and the active operating mode.

In memory mode, the proposed architecture functions as a conventional differential SRAM array. The selected word line is activated by the row decoder, allowing the stored data in the accessed cell to develop a differential voltage on BL and BLB. After sufficient bit-line separation is established, the modified sense amplifier is enabled to regenerate the small differential signal into full-swing digital outputs. Thus, the memory-mode operation follows the standard SRAM read sequence of precharge, word-line activation, bit-line development, and sense-amplifier evaluation. Since the proposed cell provides differential internal storage nodes and a controlled access path, the read operation benefits from improved signal separation and robust sensing.

In compute mode, the same array-level bit-line and sensing infrastructure is reused for logic evaluation. Instead of transferring the stored operands from the SRAM array to an external logic unit, the operands stored in selected cells interact through the shared bit-line structure. The timing-control block activates the required control signals so that the selected cells generate a logic-dependent response on BL and BLB. This response is then sensed and forwarded to the IMC logic path to generate Boolean outputs such as NAND, NOR, XOR, and XNOR. Therefore, the compute operation is performed using the stored data directly inside the SRAM array, reducing the movement of operands between memory and external logic.

The figure also highlights the manner in which the earlier objectives are combined within the final array-level design. The proposed sense amplifier from the first objective provides the high-speed and variation-tolerant readout mechanism. The proposed differential 8T SRAM cell from the second objective provides the storage element and the controlled access paths required for robust memory and compute operation. The SRAM-IMC macro and logic path developed in the third objective provide the Boolean computation capability. In the present architecture, these blocks are not treated as isolated contributions; rather, they are combined into a single memory-compute data path to realize a reconfigurable SRAM-IMC array.

The key advantage of this architecture is peripheral reuse. The bit-line pair, precharge devices, row-selection network, timing-control block, and sense amplifier are shared between memory and compute modes. This avoids the need for a fully separate compute peripheral attached to the SRAM array. As a result, the architecture retains the basic functionality of a conventional SRAM array while adding in-memory Boolean computation capability. This is important because the usefulness of SRAM-IMC depends not only on the correctness of logic operation, but also on whether the compute function can be added without degrading normal memory access.

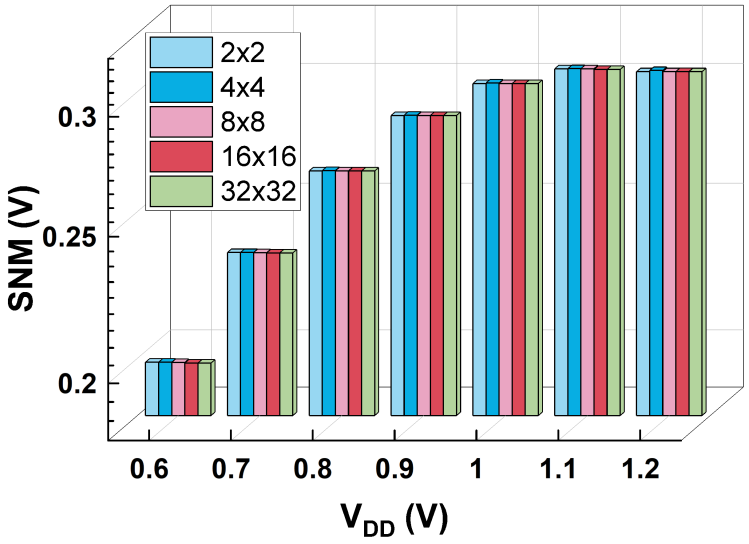
The proposed architecture therefore establishes a common framework for evaluating both memory-mode and compute-mode behavior. In memory mode, the important metrics are static noise margin, write-margin deviation, read delay, and read power. In compute mode, the important metrics are logic-evaluation latency, average energy per bit, and throughput. Since both modes share the same physical bit-line and sensing path, the array-level analysis must consider how scaling the array size affects both storage reliability and compute performance. The following sections therefore evaluate the proposed architecture from 2×2 to 32×32 array sizes across the supply-voltage range of V_{DD} = 0.6–1.2 V.

5.4 MEMORY-MODE ANALYSIS UNDER ARRAY SCALING

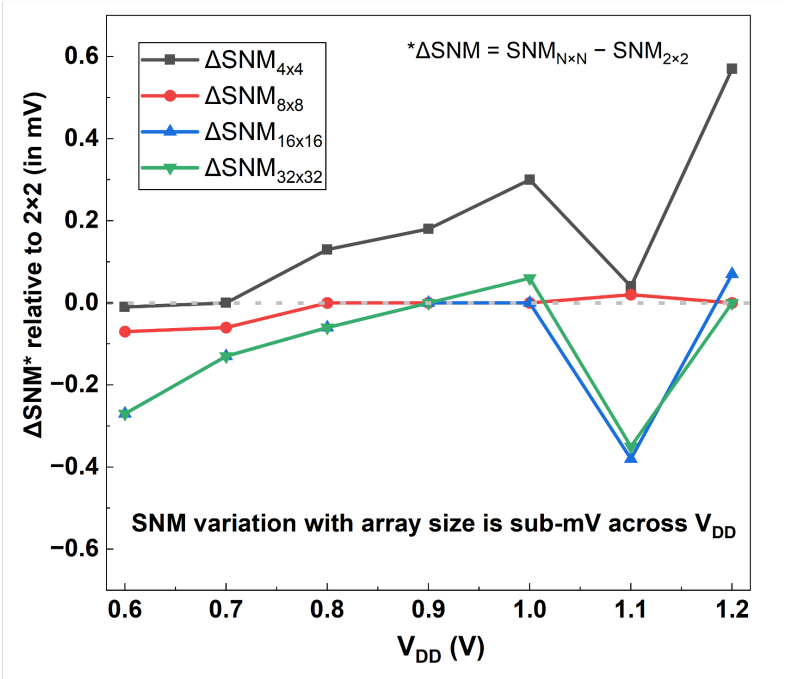
The proposed SRAM-IMC array must retain reliable conventional SRAM operation before its compute-mode functionality can be considered useful. Therefore, the first part of the array-level evaluation focuses on memory-mode behavior. The memory-mode metrics considered in this section are static noise margin (SNM), write-margin deviation, read delay, and read power. These metrics are evaluated for array sizes ranging from 2×2 to 32×32 over a supply-voltage range of V_{DD} = 0.6–1.2 V.

Array-level SRAM behavior is strongly affected by the interaction between the bit-cell, bit-line loading, word-line loading, and peripheral circuitry. Previous SRAM array studies have shown that access time, energy, and yield are governed not only by the intrinsic bit-cell design, but also by array organization, bit-line capacitance, sense-amplifier behavior, and statistical variation [2, 8, 73, 74]. In SRAM-CIM and SRAM-IMC arrays, these constraints become even more important because the same memory infrastructure is expected to support both storage and computation [7]. Hence, the memory-mode analysis in this section is used to verify that the proposed reconfigurable architecture does not sacrifice the fundamental SRAM read/write behavior when scaled to larger array sizes.

5.4.1 STATIC NOISE MARGIN



(a)



(b)

Figure 5.2: SNM scaling of the proposed SRAM cell for 2×2 to 32×32 arrays over V_{DD} = 0.6–1.2 V: (a) absolute SNM and (b) relative SNM deviation, ΔSNM with respect to the 2×2 reference array.

Figure 5.2 shows the SNM variation of the proposed SRAM cell under array-size scaling. Figure 5.2 (a) presents the absolute SNM values, while Fig. 5.2 (b) shows the relative SNM deviation with respect to the 2×2 array. The relative deviation is defined as

$$\Delta\text{SNM}_{N\times N} = \text{SNM}_{N\times N} - \text{SNM}_{2\times 2}. \tag{5.1}$$

The absolute SNM increases with V_{DD}, as expected from the stronger inverter transfer characteristics at higher supply voltage. From the plotted data, the SNM increases from approximately 0.208 V at V_{DD} = 0.6 V to approximately 0.318–0.320 V at V_{DD} = 1.2 V, corresponding to an improvement of nearly 53%. Most of this improvement occurs between 0.6 V and 1.0 V, where the SNM rises from about 0.208 V to about 0.315 V. Beyond 1.0 V, the SNM begins to saturate, and the additional improvement from 1.0 V to 1.2 V is less than 2%.

More importantly, the SNM values for all array sizes remain almost overlapped. The relative SNM variation in Fig. 5.2 (b) remains within approximately −0.38 mV to +0.57 mV across the complete voltage range. Even in the worst case, this corresponds to less than 0.3% of the absolute SNM at the lowest supply voltage. At nominal and higher supply voltages, the percentage deviation is even smaller. This confirms that increasing the array size from 2×2 to 32×32 introduces negligible degradation in static cell stability.

This behavior is consistent with the nature of SNM as a primarily local storage-node metric. Since SNM is extracted from the butterfly characteristics of the cross-coupled storage inverters, it is more strongly influenced by the bit-cell ratio, device mismatch, and supply voltage than by the external array loading. Array parasitics have a stronger effect during dynamic read and write operations, but they do not significantly alter the static hold condition when the storage nodes are properly isolated. Similar observations have been reported in array-level SRAM reliability studies, where statistical cell variation and supply scaling dominate SNM behavior, while peripheral loading primarily affects access delay and energy [2, 73]. Therefore, the proposed cell preserves its static stability when extended to larger SRAM-IMC arrays.

5.4.2 WRITE-MARGIN DEVIATION

Figure 5.3 presents the array-size dependence of the write-margin deviation. The plotted metric is $|WM - V_{DD}/2|$, which represents the distance of the write margin from the ideal midpoint value. A smaller value indicates that the write operation is closer to the desired symmetric switching condition.

Unlike SNM, the write-margin deviation shows stronger dependence on both supply voltage

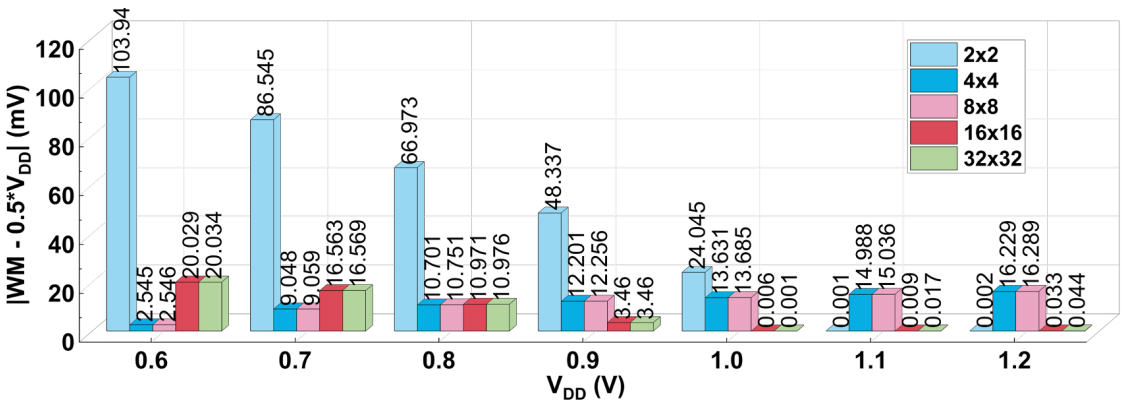


Figure 5.3: Array-size dependence of write-margin deviation, $|WM - V_{DD}/2|$, across the supply-voltage range.

and array size. At V_{DD} = 0.6 V, the 2×2 array shows the largest deviation, approximately 110 mV, indicating that write operation is most sensitive near the lowest evaluated supply voltage. As V_{DD} increases, this deviation reduces to about 70 mV at 0.8 V, about 30 mV at 1.0 V, and approaches the ideal midpoint at 1.1–1.2 V. Thus, for the 2×2 case, the write-margin deviation reduces by approximately 73% between 0.6 V and 1.0 V.

For the larger 16×16 and 32×32 arrays, the deviation is approximately 25 mV at 0.6 V, decreases to about 15 mV near 0.8 V, and further reduces to nearly 5 mV at higher supply voltages. This corresponds to an approximate reduction of 80% from the low-voltage condition to the high-voltage region. The 4×4 and 8×8 arrays remain in an intermediate range, with deviations of the order of a few millivolts at low voltage and approximately 20 mV at higher supply voltage. Overall, the plot indicates that write margin is more sensitive to array configuration than SNM, particularly below 0.8 V.

This trend is expected because write operation depends on the ability of the access devices and write drivers to overpower the existing stored state. At low V_{DD}, the reduced gate overdrive weakens the access path and makes the switching point more sensitive to device strength, control timing, and bit-line loading. Prior SRAM works have shown that writability and half-select robustness become more difficult under low-voltage and array-level operation, often motivating write-assist schemes, half-select mitigation, or modified cell structures [52, 55, 59, 73]. Therefore, while the proposed cell maintains excellent static stability under scaling, write margin remains a more informative metric for identifying low-voltage array-level limitations.

5.4.3 READ DELAY

Figure 5.4 shows the read-delay variation with array size across the supply-voltage sweep. The read delay decreases strongly as V_{DD} increases, which reflects the improvement in bit-line development and sense-amplifier regeneration at higher supply voltages.

At V_{DD} = 0.6 V, the read delay is approximately 212 ps for the 2×2 array and approximately 244 ps for the 32×32 array. This represents an array-size penalty of about 15% at the lowest supply voltage. However, the penalty reduces rapidly as the supply voltage increases. At V_{DD} = 0.8 V, the read delay is approximately 50 ps for all array sizes, and beyond 0.9 V, the difference between the smallest and largest arrays becomes only a few picoseconds. At V_{DD} = 1.2 V, the delay is close to 30–31 ps, indicating that the size-dependent read-delay penalty is nearly suppressed in the high-supply region.

The voltage-scaling trend is also significant. For the 2×2 array, the read delay reduces from about 212 ps at 0.6 V to about 30 ps at 1.2 V, corresponding to an approximate 86% reduction. For the 32×32 array, the delay decreases from about 244 ps to about 31 ps, corresponding to an approximate 87% reduction. Thus, the proposed array benefits strongly from supply-voltage scaling in memory read mode, while the effect of array size becomes small at moderate and high supply voltages.

The observed behavior agrees with the known dependence of SRAM read delay on bit-line capacitance and sense-amplifier resolution. Larger arrays increase the effective bit-line and

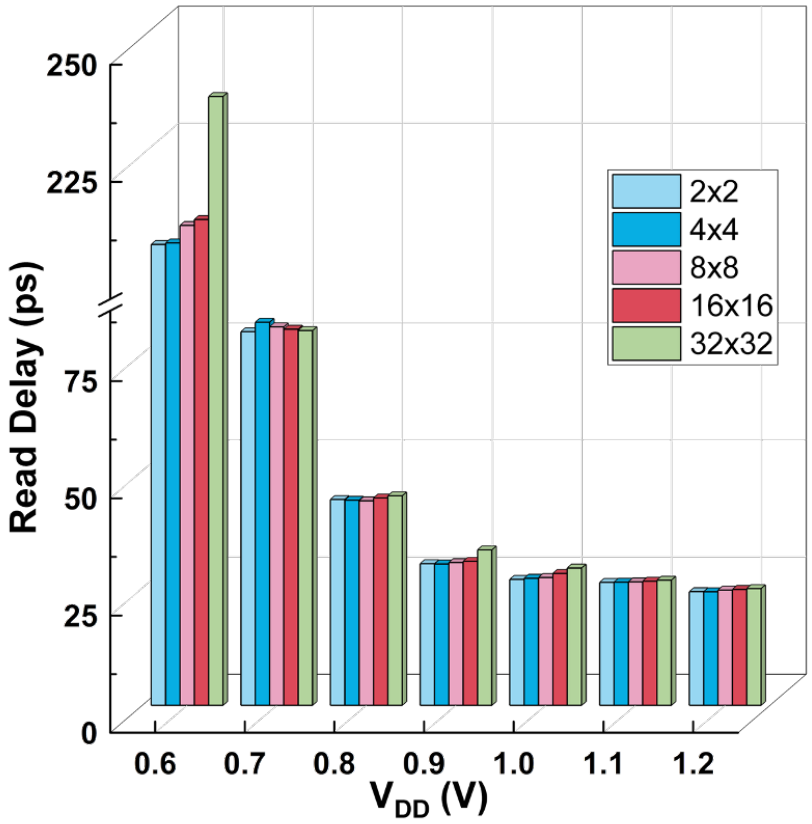


Figure 5.4: Read-delay variation with array size from 2×2 to 32×32 across the supply-voltage sweep.

interconnect loading, which slows the development of the differential signal at low voltage. However, higher V_{DD} increases the read-current drive and strengthens the regenerative action of the sense amplifier, thereby reducing the impact of the additional loading. Previous sense-amplifier and array-level SRAM studies have similarly shown that access delay is strongly linked to bit-line swing, capacitive loading, and latch resolution time [8, 30, 34]. Therefore, the proposed architecture maintains good read-speed scalability, with the main delay penalty appearing only under the combined condition of large array size and low supply voltage.

5.4.4 READ POWER

Figure 5.5 shows the read-power variation across array sizes and supply voltages. The y-axis is plotted on a logarithmic scale because the read-power spread across the evaluated array sizes is much larger than the spread observed for SNM or read delay.

Read power increases with both V_{DD} and array size. From the plotted data, the 2×2 array increases from approximately 10^{-3} μW at 0.6 V to approximately 6×10^{-2} μW at 1.2 V, corresponding to roughly a 60× increase. For the 32×32 array, the power increases from approximately 0.8 μW at 0.6 V to approximately 40 μW at 1.2 V, corresponding to roughly a 50× increase. Therefore, although the read delay improves with increasing V_{DD}, the read-power cost rises significantly.

At a fixed supply voltage, the array-size penalty is even more pronounced. At $V_{DD} = 0.6$ V,

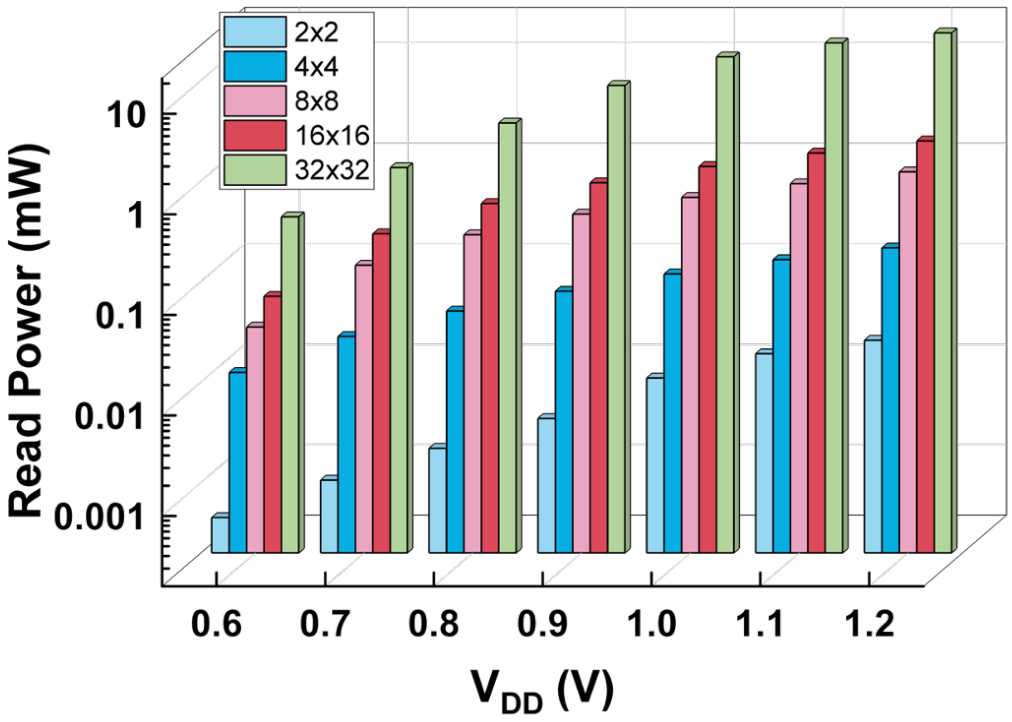


Figure 5.5: Read-power variation with array size from 2×2 to 32×32 across the supply-voltage sweep. The logarithmic y-axis highlights the wide power spread under array scaling.

the 32×32 array consumes approximately 10^2–10^3 times more read power than the 2×2 array. A similar order-of-magnitude separation is visible at higher supply voltages. This large spread shows that read power is the most strongly array-size-dependent memory-mode metric among those evaluated in this section.

This trend is physically expected because read power is directly linked to switched capacitance, bit-line length, peripheral activity, and interconnect loading. As the array grows, the bit-line capacitance and the number of capacitively coupled nodes increase, so the energy required for precharge, bit-line development, and sensing also increases. Energy-optimized SRAM array studies have shown that array organization and bit-line segmentation strongly influence read energy [8, 74]. In SRAM-CIM and SRAM-IMC architectures, peripheral and bit-line energy can become a dominant scaling limitation, especially when the same array is reused for compute-mode operation [7]. Hence, read power becomes the primary memory-mode scaling penalty in the proposed reconfigurable array.

5.4.5 INTERPRETATION OF MEMORY-MODE RESULTS

Table 5.1 summarizes the dominant memory-mode trends extracted from Figs. 5.2–5.5. The results show that the proposed SRAM-IMC array maintains strong static stability under scaling, but the dynamic metrics become increasingly sensitive to array size and supply voltage.

Table 5.1: Summary of memory-mode scaling trends for the proposed reconfigurable SRAM-IMC array.

Metric	Observed trend	Scaling implication
SNM	Increases from ∼ 0.208 V at 0.6 V to ∼ 0.32 V at 1.2 V, while ΔSNM remains within about ±0.6 mV.	Negligible array-size impact; stability remains well preserved.
Write-margin deviation	Worst low-voltage deviation is ∼ 110 mV for 2×2 at 0.6 V, reducing by about 73% by 1.0 V. Larger arrays show lower high-V_{DD} deviations of only a few millivolts.	Moderate scaling sensitivity, strongest below 0.8 V.
Read delay	For 2×2, delay reduces from ∼ 212 ps to ∼ 30 ps; for 32×32, it reduces from ∼ 244 ps to ∼ 31 ps. The low-V_{DD} size penalty is about 15%.	Moderate size penalty at low voltage; small penalty at nominal/high voltage.
Read power	Increases strongly with both V_{DD} and array size. The 32×32 array is roughly 10^2–10^3 times higher than the 2×2 case across the sweep.	Dominant memory-mode scaling cost due to switched capacitance and peripheral loading.

Overall, the memory-mode analysis shows that array scaling has negligible impact on the static stability of the proposed cell. The maximum SNM shift relative to the 2×2 reference remains below 0.6 mV, which is less than 0.3% of the lowest absolute SNM value observed. Write-margin deviation and read delay show moderate scaling sensitivity, especially at low V_{DD}, where reduced device drive makes the array more sensitive to loading and control timing. Read power exhibits the strongest dependence on array size because it scales directly with switched capacitance and peripheral activity.

Therefore, the proposed reconfigurable SRAM-IMC array preserves reliable memory-mode behavior up to the evaluated 32×32 size. The main limitation introduced by array scaling is not static cell stability, but dynamic power consumption. This conclusion is important for the compute-mode analysis in the next section, because the same bit-line and sensing resources are reused for in-memory Boolean logic evaluation.

5.5 COMPUTE-MODE ANALYSIS UNDER ARRAY SCALING

The preceding section confirmed that the proposed reconfigurable SRAM-IMC array preserves reliable memory-mode behavior under array-size scaling. The next step is to evaluate whether the same array can also sustain efficient compute-mode operation. In compute mode, the stored operands are not first read out and then processed by a separate external logic block. Instead, the

selected SRAM cells generate a logic-dependent bit-line response that is sensed and interpreted by the IMC data path. Therefore, the compute-mode delay and energy are governed not only by the intrinsic switching behavior of the bit cell, but also by the bit-line capacitance, interconnect loading, precharge activity, timing-control sequence, and sense-amplifier response.

The compute-mode analysis in this section considers NAND evaluation for the input case $A = 0$ and $B = 0$. This case is used as a representative logic-evaluation condition because it exercises the complete SRAM-IMC compute path, including cell selection, bit-line development, sensing, and output generation. The array sizes considered are 2×2, 4×4, 8×8, 16×16, and 32×32, and the supply voltage is swept from $V_{DD} = 0.6$ V to 1.2 V. The key compute-mode metrics are logic-evaluation latency, average energy per bit, and throughput.

Prior SRAM-CIM and SRAM-IMC works have shown that compute performance at the array level is frequently limited by the same factors that affect memory readout: bit-line loading, sensing margin, peripheral timing, and control overhead [84, 88, 106]. More recent SRAM-CIM designs have further emphasized that peripheral circuitry and bit-line energy can dominate the overall energy efficiency, particularly when the array supports configurable precision, multi-bit operation, or mixed memory/compute functionality [7, 82, 104]. Hence, the purpose of this section is not only to verify that the proposed SRAM-IMC array computes correctly, but also to identify how the latency, energy, and throughput scale when the macro is extended to larger array organizations.

5.5.1 COMPUTE-MODE LATENCY

Figure 5.6 shows the compute-mode latency of the proposed SRAM-IMC array as a function of V_{DD} and array size. The plotted delay corresponds to the NAND-evaluation delay for $A = 0$ and $B = 0$.

The compute latency decreases as V_{DD} increases for all array sizes. From the plotted data, the latency at $V_{DD} = 0.6$ V lies approximately in the range of 493–517 ps, whereas at $V_{DD} = 1.2$ V, it reduces to approximately 380–388 ps. Thus, the compute delay improves by nearly 23%–25% across the supply-voltage sweep. This reduction is expected because a higher supply voltage increases the overdrive of the access and evaluation transistors, accelerates bit-line voltage development, and improves the regeneration speed of the sensing path.

The array-size dependence is most visible at the lowest supply voltage. At 0.6 V, the difference between the fastest and slowest array cases is approximately 20–25 ps, corresponding to an array-size-induced spread of about 4%–5%. As V_{DD} increases, this spread becomes smaller. At 1.2 V, the latency spread reduces to approximately 8–10 ps, or about 2%–3%. Therefore, while larger arrays do introduce additional loading, the delay penalty does not grow proportionally with array size.

This trend indicates that the proposed compute path remains sensing-limited rather than severely load-limited up to the evaluated 32×32 array size. In other words, the additional bit-line and interconnect capacitance of the larger arrays affects the low-voltage case, but the sensing path still resolves the compute output within a narrow delay window. This observation is

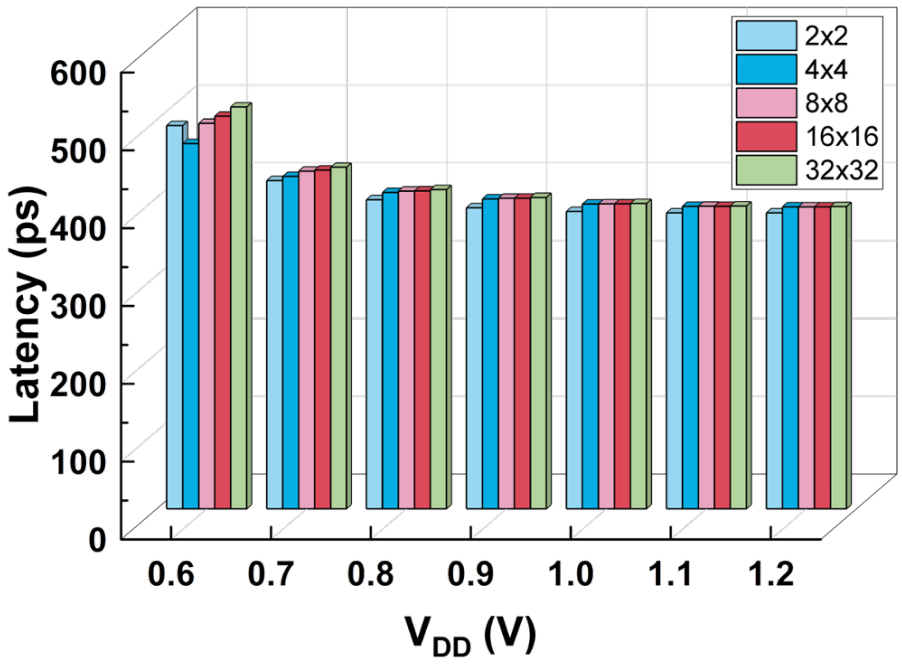


Figure 5.6: Compute-mode NAND-evaluation latency of the proposed SRAM-IMC array for 2×2 to 32×32 organizations across $V_{DD} = 0.6$–1.2 V.

consistent with prior SRAM-CIM analyses, where higher V_{DD} reduces the sensitivity of bit-line evaluation to parasitic loading, while low-voltage operation exposes the delay penalty associated with weaker current drive and larger effective capacitance [7, 106]. Consequently, the proposed SRAM-IMC architecture shows favorable latency scalability, especially at nominal and higher supply voltages.

The latency behavior can be interpreted as the sum of three main components:

$$t_{comp} = t_{BL} + t_{SA} + t_{logic}, \tag{5.2}$$

where t_{BL} is the time required to develop the compute-dependent bit-line response, t_{SA} is the sense-amplifier resolution delay, and t_{logic} is the output logic delay. Increasing the array size mainly increases t_{BL} through additional capacitance, whereas increasing V_{DD} reduces both t_{BL} and t_{SA}. This explains why the array-size penalty is more visible at 0.6 V, but becomes small as the supply voltage approaches 1.0–1.2 V.

5.5.2 AVERAGE COMPUTE ENERGY PER BIT

Figure 5.7 shows the average energy per bit consumed during compute-mode NAND evaluation. The y-axis is plotted on a logarithmic scale because the energy spread across array sizes is much larger than the corresponding latency spread.

The average compute energy per bit increases with both supply voltage and array size. For the 2×2 array, the energy is approximately 0.008–0.01 nJ at 0.6 V, and it increases to approximately

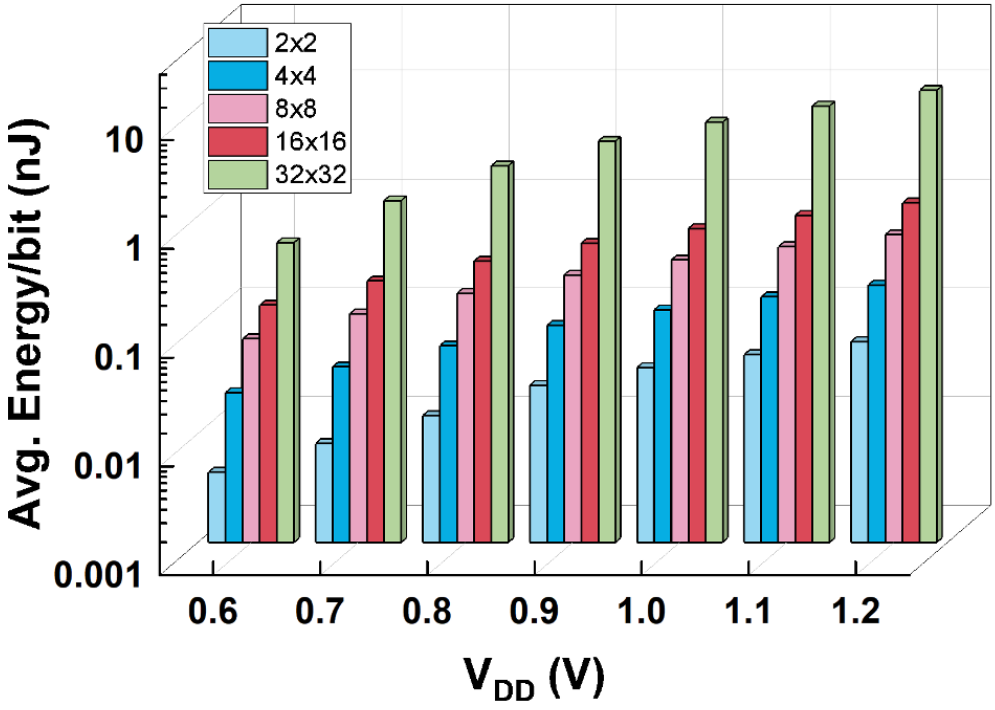


Figure 5.7: Average compute energy per bit of the proposed SRAM-IMC array for 2×2 to 32×32 organizations across $V_{DD} = 0.6$–1.2 V.

0.13–0.15 nJ at 1.2 V. This corresponds to an increase of roughly 15×–18×. For the 32×32 array, the energy increases from approximately 1 nJ at 0.6 V to approximately 25–30 nJ at 1.2 V, corresponding to an increase of about 25×–30×.

The array-size dependence is even more significant than the voltage dependence. At $V_{DD} = 0.6$ V, the 32×32 array consumes roughly two orders of magnitude more energy per bit than the 2×2 array. A similar separation is observed at higher supply voltages; at 1.2 V, the 32×32 energy is approximately 150×–200× larger than that of the 2×2 array. This shows that compute energy is the dominant scaling penalty of the proposed architecture.

The physical reason for this behavior is the increase in effective switched capacitance. During compute evaluation, energy is consumed in precharging, conditionally discharging, and sensing the bit-line and compute-path nodes. The dynamic energy can be approximated by

$$E_{dyn} \approx \alpha C_{eff} V_{DD}^2, \tag{5.3}$$

where α is the activity factor and C_{eff} is the effective switched capacitance seen by the compute path. Increasing V_{DD} raises the energy approximately with V_{DD}^2, while increasing the array size raises C_{eff} through longer bit-lines, larger interconnect networks, and increased peripheral loading. Therefore, even though the latency penalty remains modest, the compute-energy penalty becomes large.

This observation is aligned with the broader SRAM-CIM literature. Several SRAM-

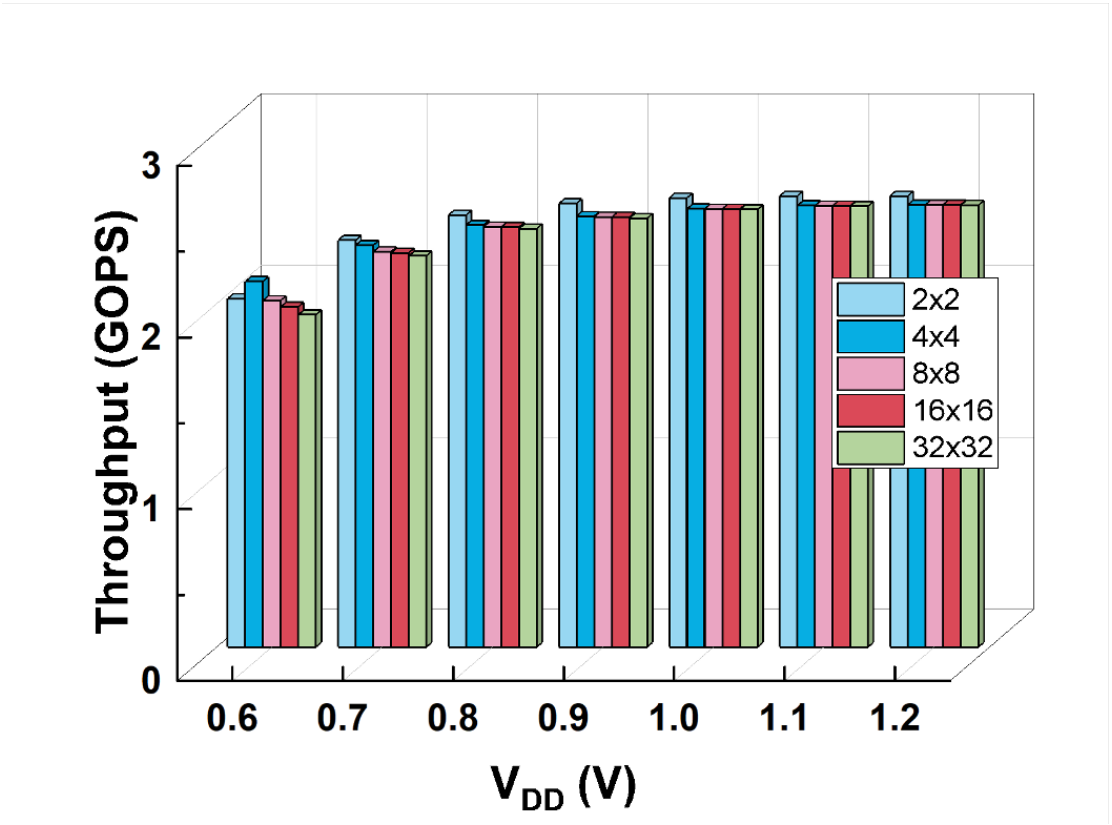


Figure 5.8: Compute-mode throughput of the proposed SRAM-IMC array for 2×2 to 32×32 organizations across V_{DD} = 0.6–1.2 V.

CIM macros achieve high throughput by exploiting column-parallelism and internal bit-line bandwidth, but the total energy efficiency is often determined by peripheral switching, bit-line capacitance, and sensing/ADC overhead rather than by the Boolean or arithmetic operation alone [82, 84, 88, 104]. In the proposed design, the same principle is observed at the Boolean-logic level: the compute operation remains fast, but the energy cost increases substantially as more capacitance is associated with the array-level data path. Thus, energy per bit is the most important compute-mode metric for future optimization.

5.5.3 COMPUTE THROUGHPUT

Figure 5.8 presents the compute-mode throughput of the proposed SRAM-IMC array. Since the plotted operation is a single-bit NAND evaluation through the SRAM-IMC path, the throughput follows the inverse trend of compute latency.

The throughput increases with V_{DD} for all array sizes. From the plotted data, the throughput is approximately 2.1–2.3 GOPS at 0.6 V, and it increases to approximately 2.75–2.85 GOPS at 1.2 V. This corresponds to an improvement of about 25%–32%, depending on the array size. The trend is consistent with Fig. 5.6, since reducing the compute delay directly increases the number of logic evaluations that can be completed per second.

The throughput is defined as

$$\text{Throughput} = \frac{1}{t_{\text{comp}}}, \tag{5.4}$$

where t_{comp} is the compute-mode latency. Therefore, the observed improvement in throughput is a direct consequence of the reduction in bit-line evaluation and sensing delay at higher supply voltage. For example, reducing the latency from about 500 ps to about 385 ps corresponds to an increase from about 2.0 GOPS to about 2.6 GOPS, which is consistent with the measured throughput range.

The array-size dependence of throughput is comparatively small. At 0.6 V, the spread between the smallest and largest array sizes is approximately 0.2–0.25 GOPS, or around 9%–11%. At 1.2 V, the spread reduces to less than 0.05 GOPS, corresponding to below about 2%. This confirms that the proposed array does not suffer a large throughput penalty when scaled from 2×2 to 32×32, particularly in the nominal and high-supply regions.

This result is significant because it separates the speed-scaling behavior from the energy-scaling behavior. The proposed SRAM-IMC array remains throughput-stable under array-size scaling, but the energy per bit increases strongly. Similar speed–energy separation has been observed in SRAM-CIM macros, where bit-line parallelism supports high compute throughput, while peripheral and capacitive switching determine the energy cost [7, 84, 88, 106]. Therefore, the proposed architecture is promising from a compute-throughput perspective, but energy-aware array partitioning and peripheral optimization are necessary for larger implementations.

5.5.4 INTERPRETATION OF COMPUTE-MODE RESULTS

Table 5.2 summarizes the compute-mode trends extracted from Figs. 5.6–5.8. The results show that the proposed reconfigurable SRAM-IMC array scales well in terms of latency and throughput, while energy per bit becomes the dominant limitation as the array size increases.

Overall, the compute-mode analysis demonstrates that the proposed SRAM-IMC array maintains stable logic-evaluation behavior when scaled from 2×2 to 32×32. The latency decreases by approximately one-fourth across the voltage sweep, and the throughput improves correspondingly. The array-size penalty in latency and throughput is modest, especially above 0.9 V, where the bars for different array sizes become closely grouped. This indicates that the shared bit-line and sensing infrastructure can support compute-mode operation without a major degradation in speed.

In contrast, the average compute energy per bit increases substantially with array size. This is because array expansion increases the effective switched capacitance of the compute data path, including the bit-lines, interconnects, precharge devices, and peripheral sensing circuits. Therefore, the principal trade-off of the proposed reconfigurable SRAM-IMC array is not compute speed, but compute energy. This conclusion is consistent with the memory-mode analysis, where SNM and read delay remained well controlled, while read power increased strongly with array size.

The combined memory-mode and compute-mode results show that the proposed architecture remains functionally robust and throughput-efficient up to the evaluated 32×32 array size. However, the energy trend indicates that future scaling should focus on reducing capacitance-related overhead. Possible directions include bit-line segmentation, local precharge gating, selective

Table 5.2: Summary of compute-mode scaling trends for the proposed reconfigurable SRAM-IMC array.

Metric	Observed trend	Design implication
Compute latency	Reduces from approximately 493–517 ps at 0.6 V to approximately 380–388 ps at 1.2 V, giving nearly 23%–25% delay reduction. The array-size spread reduces from about 4%–5% at low V_{DD} to about 2%–3% at high V_{DD}.	Latency scales favorably; size penalty is mainly visible at low voltage.
Average energy per bit	Increases strongly with both V_{DD} and array size. The 32×32 array is roughly 10^2 times more energy-consuming than the 2×2 array, and the energy of the 32×32 case rises by about 25×–30× from 0.6 V to 1.2 V.	Energy is the dominant compute-mode scaling penalty.
Throughput	Increases from approximately 2.1–2.3 GOPS at 0.6 V to approximately 2.75–2.85 GOPS at 1.2 V. The array-size spread falls below about 2% at high V_{DD}.	Throughput remains stable across array sizes, especially near nominal supply.

row/column activation, reduced-swing compute signaling, and more aggressive peripheral sharing. These optimizations would directly target the dominant energy component while preserving the latency and throughput advantages demonstrated by the proposed SRAM-IMC data path.

5.6 ARRAY-LEVEL SCALABILITY AND DESIGN TRADE-OFFS

The memory-mode and compute-mode analyses together show that the proposed reconfigurable SRAM-IMC architecture can be extended from small arrays to larger organizations without losing its basic memory or compute functionality. However, the different performance metrics do not scale in the same manner. Static stability, read delay, compute latency, and throughput remain relatively well controlled as the array size increases, whereas read power and compute energy show a much stronger dependence on array dimension. This indicates that the dominant limitation of the proposed array is not functional correctness or timing scalability, but energy growth caused by larger switched capacitance.

In memory mode, the most favorable result is the weak dependence of SNM on array size. The relative SNM deviation with respect to the 2×2 reference remains within the sub-mV range across the complete supply-voltage sweep. This confirms that the intrinsic storage stability of the proposed bit cell is preserved even when the array is scaled to 32×32. The result is important because it shows that the proposed differential 8T cell does not suffer a significant static-stability penalty when embedded in a larger memory structure. Since SNM is primarily governed by the

local cross-coupled storage nodes and supply voltage, the external array parasitics have only limited influence on the hold condition.

Write-margin deviation and read delay show a stronger dependence on supply voltage and array size. At low V_{DD}, the reduced device drive makes the cell more sensitive to bit-line loading and control-signal timing. Therefore, larger arrays exhibit a more noticeable penalty in write and read behavior near 0.6 V. As the supply voltage increases, the stronger access and sensing currents reduce this penalty, and the delay spread across array sizes becomes smaller. This shows that the proposed architecture is more suitable for nominal-voltage operation when larger arrays are used, while low-voltage operation requires more careful timing and write-assist consideration.

The strongest memory-mode scaling penalty appears in read power. Unlike SNM, read power is directly related to the amount of capacitance that is charged, discharged, and sensed during access. As the array grows, longer bit-lines, larger interconnect networks, and increased peripheral loading raise the effective switched capacitance. This causes the read power to increase by orders of magnitude from the smallest to the largest evaluated array. Thus, while the proposed bit cell remains stable under scaling, the energy cost of accessing larger arrays becomes the main memory-mode concern. This observation is consistent with SRAM array studies where bit-line and peripheral capacitance dominate energy consumption in scaled memory organizations [8, 74].

A similar trend is observed in compute mode. The NAND-evaluation latency decreases with increasing V_{DD}, and the array-size-induced delay spread remains limited. From the plotted data, the compute latency reduces from approximately 493–517 ps at 0.6 V to approximately 380–388 ps at 1.2 V. This corresponds to a delay reduction of nearly 23%–25%. The array-size penalty is most visible at low supply voltage, but it becomes small in the nominal and high-V_{DD} regions. Therefore, the shared bit-line and sensing infrastructure can support compute-mode operation without introducing a large delay penalty up to the evaluated 32×32 array size.

Throughput follows the same trend as compute latency. Since throughput is inversely related to the compute-evaluation time, it increases as V_{DD} increases. The throughput values for different array sizes remain closely grouped at higher supply voltages, indicating that array expansion does not severely reduce compute-rate capability. This is a favorable result for the proposed SRAM-IMC architecture because the objective of the design is throughput enhancement through in-array Boolean evaluation. The results show that the compute data path remains fast enough to preserve this advantage under moderate array scaling.

However, the average compute energy per bit increases strongly with both V_{DD} and array size. This is the dominant compute-mode trade-off. Although larger arrays do not significantly degrade latency or throughput, they require substantially more energy because the compute operation charges, discharges, and senses larger bit-line and peripheral capacitances. Therefore, the proposed architecture is speed-scalable but energy-limited. This behavior is consistent with recent SRAM-CIM literature, where bit-line switching, sensing circuitry, and peripheral overhead are often the main contributors to total energy even when the computational operation itself is simple [7, 84, 88, 104].

Table 5.3: Overall array-level scaling behavior of the proposed reconfigurable SRAM-IMC architecture.

Operating mode	**Metric**	**Scaling behavior and implication**
Memory mode	SNM	Weakly affected by array size; relative shift remains within the sub-mV range, confirming stable storage behavior up to 32×32.
Memory mode	Write-margin deviation	Moderately affected by array size and low-V_{DD} operation; indicates that writability is more sensitive to scaling than hold stability.
Memory mode	Read delay	Moderate low-voltage size penalty; delay spread reduces at higher V_{DD}, showing good read-speed scalability near nominal supply.
Memory mode	Read power	Strongest memory-mode scaling penalty; increases with array size because of larger switched bit-line and peripheral capacitance.
Compute mode	Latency	Reduces by nearly 23%–25% across the voltage sweep; array-size penalty remains limited, especially above 0.9 V.
Compute mode	Average energy per bit	Dominant compute-mode scaling penalty; increases strongly with both V_{DD} and array size due to higher effective switched capacitance.
Compute mode	Throughput	Improves with V_{DD} and remains closely grouped across array sizes at higher supply voltages, indicating stable compute-rate scalability.

The combined results suggest that the proposed architecture is most effective when the design goal is to improve throughput by reducing data movement while maintaining conventional SRAM functionality. The shared data path allows the same array to operate in memory and compute modes, avoiding the need for a completely separate compute peripheral. This improves architectural compactness and supports local Boolean evaluation. However, because both modes

rely on the same physical bit-line infrastructure, energy optimization becomes increasingly important as the array grows.

Table 5.3 summarizes the overall scaling behavior observed in the proposed reconfigurable SRAM-IMC array.

From the design perspective, these trends indicate that future improvements should primarily target capacitance-related energy overhead rather than the basic logic-evaluation mechanism. Possible optimization strategies include bit-line segmentation, selective row and column activation, local precharge gating, reduced-swing compute operation, and more aggressive sharing of peripheral circuits. These techniques would reduce the switched capacitance seen during both read and compute operations, directly addressing the strongest scaling penalty observed in this chapter.

Overall, the array-level analysis confirms that the proposed SRAM-IMC design preserves the key benefits established in the previous chapters. The modified sensing strategy supports fast differential readout, the proposed 8T cell maintains stable memory behavior, and the IMC macro enables in-array Boolean logic evaluation. When these blocks are integrated into one reconfigurable array, the architecture remains stable and throughput-efficient up to the evaluated 32×32 size. The main limitation is the increase in energy and power under array expansion, which identifies energy-aware peripheral and bit-line optimization as the most important direction for further scaling.

5.7 SUMMARY OF IMPORTANT RESULTS

This chapter integrated the proposed sense amplifier, differential 8T SRAM cell, and SRAM-IMC logic path into a reconfigurable array-level architecture. The array was evaluated in both memory and compute modes for sizes ranging from 2×2 to 32×32 across $V_{DD} = 0.6$–1.2 V. The important results obtained from this analysis are summarized below.

1. The proposed architecture establishes a unified dual-mode SRAM-IMC array in which the same bit-line and sensing infrastructure is used for both conventional memory operation and in-memory Boolean computation. In memory mode, the selected cell develops a differential BL/BLB signal that is amplified by the sense amplifier. In compute mode, the sensed bit-line response is forwarded to the IMC logic path to generate outputs such as NAND, NOR, XOR, and XNOR.

2. The static noise margin remains nearly unaffected by array scaling. From 2×2 to 32×32 arrays, the relative SNM variation with respect to the 2×2 reference remains within approximately ± 0.6 mV over the complete supply-voltage range. This confirms that the proposed bit cell preserves its static storage stability even when embedded in larger array organizations.

3. The write-margin deviation, expressed as $|WM - V_{DD}/2|$, shows a stronger dependence on supply voltage and array size than SNM. The largest deviations occur at lower supply voltages, especially below 0.8 V, where reduced device drive makes the write operation

more sensitive to bit-line loading and control timing. As V_{DD} increases, the deviation compresses and the write behavior approaches the ideal midpoint more closely.

4. The read delay decreases strongly with increasing V_{DD} for all array sizes. The array-size penalty is most visible at low supply voltage because larger arrays introduce heavier bit-line and interconnect loading. However, beyond approximately 0.9–1.0 V, the delay values for different array sizes become closely grouped, indicating good read-speed scalability in the nominal operating region.

5. The read power shows the strongest memory-mode dependence on array size. Unlike SNM, which is mainly governed by the local storage-node stability of the cell, read power depends directly on switched capacitance, bit-line length, interconnect loading, and peripheral activity.

6. The compute-mode NAND-evaluation latency for $A = 0$ and $B = 0$ decreases from approximately 493–517 ps at $V_{DD} = 0.6$ V to approximately 380–388 ps at $V_{DD} = 1.2$ V. This represents a delay reduction of nearly 23%–25% across the supply-voltage sweep. The array-size penalty remains limited, especially at higher supply voltages, showing that compute-mode speed is not severely degraded by scaling up to 32×32.

7. The average compute energy per bit increases significantly with both supply voltage and array size. This increase is caused by the larger effective switched capacitance associated with longer bit-lines, larger interconnect networks, and additional peripheral loading in larger arrays. Thus, although compute latency remains well controlled, compute energy becomes the dominant compute-mode scaling penalty.

8. The compute throughput increases with V_{DD}, following the reduction in compute-mode latency. The throughput values for different array sizes remain closely grouped at higher supply voltages, indicating that array expansion does not introduce a large throughput penalty.

9. The combined memory-mode and compute-mode results show that the proposed SRAM-IMC array remains functionally robust up to the evaluated 32×32 size. Static stability is preserved, read and compute delays remain controlled, and throughput remains stable. The dominant trade-off introduced by scaling is the increase in read power and compute energy, both of which are linked to larger switched capacitance and peripheral activity.

Overall, the proposed reconfigurable SRAM-IMC array successfully extends the circuit-level improvements of the earlier chapters to an array-level memory-compute platform. The results show that the architecture is stable and throughput-efficient under scaling, with energy reduction identified as the main direction for further optimization.

CHAPTER – 6

CONCLUSION, FUTURE SCOPE AND SOCIAL IMPACT

The present research work is focused on SRAM design for throughput enhancement using a circuit-to-array design approach. The proposed work begins with variation-tolerant sense-amplifier design, extends to robust SRAM bit-cell development, and then progresses toward SRAM-based in-memory computation at the macro and array levels. In this chapter, the important findings and outcomes of the proposed work are summarized. The possible future extensions of the work are then discussed, followed by the potential social impact of the proposed SRAM-IMC design approach.

6.1 CONCLUSION

The present research work focused on SRAM design for throughput enhancement using a circuit-to-array design approach. The work was motivated by the increasing cost of data movement in conventional memory-centric systems and by the growing interest in SRAM-based in-memory computing, where selected logic operations are performed within or near the memory array. The thesis therefore did not treat SRAM-IMC only as a macro-level logic problem. Instead, it examined the complete design path consisting of the sense amplifier, SRAM bit-cell, SRAM-IMC macro, and reconfigurable SRAM-IMC array. This approach was necessary because the reliability and speed of SRAM-IMC depend not only on the compute logic but also on bit-line development, sensing margin, storage-node stability, and array-level capacitance.

The first important outcome of this thesis is that the sense amplifier remains a critical block for both conventional SRAM readout and SRAM-IMC operation. In scaled SRAM arrays, the bit-line voltage difference available before sensing is small, and the read decision is affected by bit-line capacitance, supply voltage, and device mismatch. To address this issue, modified decoupled sense amplifier structures, namely MDSA and MDSANF, were proposed. Both designs decouple the regenerative latch from the bit-lines during precharge and allow the latch to evaluate only after the input differential is available. This reduces unnecessary bit-line disturbance and improves sensing speed under low-voltage differential read conditions.

At V_{DD} = 0.8 V and C_{BL} = 50 fF, the MDSA achieved read-delay reductions of 28.6%, 41.79%, 37.74%, and 30.94% over MCSA, DTSA, MHSA, and LSA, respectively. The MDSANF provided comparable reductions of 26.13%, 39.78%, 35.58%, and 28.55% over the same reference topologies. Monte Carlo analysis with 10,000 samples and ±30 mV threshold-voltage variation confirmed improved variation tolerance, while sizing analysis identified an 88 nm decoupling PMOS width as a suitable power–delay design point. The layout areas of 0.297 μm^2 for MDSA and 0.5192 μm^2 for MDSANF further indicate that the speed improvement is obtained with practical column-level overhead. Thus, the proposed sensing approach establishes a faster and more variation-tolerant read path for differential SRAM and SRAM-IMC designs.

The second important outcome is that SRAM bit-cell design must be evaluated not only for storage stability but also for its suitability under array and compute-mode conditions. The proposed 8TS SRAM bit-cell was designed as a half-select-free single-ended cell in 32 nm CMOS. It achieved an SNM of 0.232 V at V_{DD} = 0.8 V, while limiting layout overhead to 4.86% relative to the 7T baseline. Monte Carlo analysis confirmed robust behavior for SNM, write margin, and access delay under threshold-voltage variation, supply fluctuation, and temperature variation. The cell also achieved $WM1$ = 0.515 V, read power of 0.186 μW, and write powers of 0.247 μW and 8.89 μW for write-1 and write-0 operations, respectively. These results show that the 8TS cell is suitable when half-select resilience and energy-aware operation are important.

In addition to the single-ended 8TS cell, the differential 8TD cell was analyzed to study the benefit of differential operation under low-voltage and variation-sensitive conditions. At V_{DD} = 0.6 V, Monte Carlo analysis showed that the 8TD cell achieved a mean SNM of approximately 0.1932 V with a standard deviation of 0.0056 V, whereas the 8TS cell achieved a lower mean SNM of approximately 0.1525 V with a higher standard deviation of 0.0152 V. This indicates that the differential cell provides better noise immunity and tighter variation spread. The conclusion from the bit-cell study is therefore not that one cell is universally best, but that different cell structures serve different design goals: 8TS is attractive for half-select-free and energy-aware storage, while 8TD is better suited for robust differential sensing and low-voltage SRAM-IMC operation.

The third important outcome is that the proposed 8TD SRAM-IMC macro demonstrates reliable bitwise Boolean computation using stored operands and a modified sensing path. In the proposed 8TD macro, two SRAM cells store operands A and B, and the shared bit-line/sensing response is used to generate NAND/NOR and XOR/XNOR logic outputs. This avoids moving the operands to an external logic unit and follows the main SRAM-IMC principle of using the memory access path itself for computation. Functional simulations verified correct operation for all input combinations, AB = 00, 01, 10, and 11, under both concurrent and sequential word-line activation schemes.

Monte Carlo simulations confirmed that the proposed 8TD SRAM-IMC macro remains reliable under device mismatch. For NAND/NOR operation, 6000 Monte Carlo runs were performed at V_{DD} = 0.8 V, T = 300.15 K, and ΔV_{TH} = 50 mV. The outputs remained separated for all input cases, the worst-case mean delay remained close to 0.4 ns, and the coefficient of variation stayed below approximately 2.5%. For XOR/XNOR operation, 2000 Monte Carlo runs gave mean delays of 0.55450 ns and 0.55425 ns, respectively, with standard deviations close to 0.0165 ns. The mean-delay difference of only about 0.25 ps indicates that the complementary XOR/XNOR paths are well balanced. This is important because SRAM-IMC correctness depends not only on average delay but also on the separation and consistency of operand-dependent bit-line responses.

The SRAM-IMC cell comparison further shows that different cell structures occupy different positions in the speed–energy design space. At V_{DD} = 0.8 V, the 8TE cell achieved 381.097 ps latency and 2.624 GOPS throughput, making it suitable for speed-oriented IMC operation. The

8TC cell achieved the lowest average energy per bit of 1.647 fJ, but with the largest latency penalty. The proposed 8TD cell reduced the mean delay relative to 8TC from 0.57549 ns to 0.55495 ns, while providing energy reduction of about 60%–70% at V_{DD} = 0.5–0.7 V, about 30%–40% near nominal supply, and about 15%–25% at higher supply voltages with respect to the 6T, 8TB, and 8TE reference cells. Therefore, the macro-level conclusion is that 8TE is more suitable for latency- and throughput-oriented IMC, whereas the proposed 8TD cell is more suitable when energy reduction is the primary design target.

The fourth important outcome is that the proposed reconfigurable SRAM-IMC array preserves the benefits of the earlier circuit-level blocks when extended to array-level operation. The architecture uses the same bit-line and sensing infrastructure for memory-mode readout and compute-mode Boolean evaluation. In memory mode, the selected cell develops a differential BL/BLB signal that is amplified by the sense amplifier. In compute mode, the sensed response is forwarded to the IMC logic path to generate outputs such as NAND, NOR, XOR, and XNOR. This dual-mode organization is significant because it avoids a fully separate compute datapath and allows the same SRAM structure to support both storage and local computation.

The memory-mode analysis from 2×2 to 32×32 arrays over V_{DD} = 0.6–1.2 V shows that static storage stability is preserved under scaling. The relative SNM deviation with respect to the 2×2 reference remains within approximately ±0.6 mV, showing that the intrinsic storage behavior of the proposed cell is not strongly disturbed by the evaluated array sizes. Write-margin deviation and read delay show greater dependence on low-voltage operation and array loading, but the delay spread becomes closely grouped near nominal supply. In contrast, read power increases strongly with array size because of larger switched bit-line and peripheral capacitance. Therefore, in memory mode, the main scaling cost is not static stability but dynamic power.

The compute-mode analysis leads to an even clearer design inference. NAND-evaluation latency reduces from approximately 493–517 ps at V_{DD} = 0.6 V to approximately 380–388 ps at V_{DD} = 1.2 V, giving nearly 23%–25% delay reduction. Throughput increases from approximately 2.1–2.3 GOPS to approximately 2.75–2.85 GOPS, corresponding to about 25%–32% improvement depending on array size. The throughput spread also becomes small at higher supply voltage, confirming that the proposed array does not suffer a large speed penalty when scaled up to 32×32. However, average compute energy per bit increases strongly with both V_{DD} and array size. The 32×32 array is roughly 10^2 times more energy-consuming than the 2×2 array, and its energy increases by about 25×–30× from 0.6 V to 1.2 V.

Thus, the most important array-level conclusion is the separation between speed scaling and energy scaling. The proposed SRAM-IMC array remains functionally robust and throughput-efficient up to the evaluated 32×32 size, but the energy per bit increases substantially because of the larger effective switched capacitance of the bit-lines, interconnects, precharge devices, and peripheral circuits. This finding is consistent with the broader understanding of SRAM-CIM systems, where internal memory parallelism can support high throughput, but the final energy cost is often dominated by bit-line and peripheral activity rather than by the Boolean operation alone. Therefore, future improvement of the proposed array should focus primarily on capacitance-aware

energy reduction rather than on changing the basic logic-evaluation mechanism.

Overall, this thesis establishes that SRAM throughput enhancement cannot be achieved by optimizing only one block. Faster sensing improves read and compute decision time, but it must be supported by a robust bit-cell. A robust bit-cell improves storage behavior, but it must also support safe compute-mode access. A correct SRAM-IMC macro demonstrates local Boolean logic, but its usefulness depends on whether the same operation remains stable and energy-efficient when placed in an array. The proposed work connects these levels through a common differential bit-line and sensing framework. The MDSA/MDSANF improve the read decision, the 8TS and 8TD cells improve robustness and application-specific operation, the proposed 8TD SRAM-IMC macro enables reliable NAND/NOR and XOR/XNOR logic, and the reconfigurable array demonstrates dual-mode memory and compute operation.

To conclude, the present research work confirms that SRAM-IMC design is a co-design problem involving bit-cell stability, bit-line behavior, sense-amplifier timing, and array-level capacitance. The proposed circuits and architectures provide a practical path for low-voltage SRAM readout and local Boolean computation. At the same time, the results show that the optimized choice depends on the target design requirement. The 8TE structure is preferable when speed and throughput are dominant, the 8TS cell is preferable for energy-aware and half-select-resilient storage, and the proposed 8TD cell provides a stronger energy-oriented SRAM-IMC option while supporting robust differential SRAM-IMC operation. The proposed reconfigurable array is most promising when local Boolean computation is required without sacrificing conventional SRAM functionality. The main limitation identified by the thesis is the growth of read power and compute energy under array scaling, which defines the most important direction for future SRAM-IMC optimization.

6.2 FUTURE SCOPE

The present work addresses SRAM design for throughput enhancement through circuit-level and array-level simulation of sense amplifiers, SRAM bit-cells, SRAM-IMC macros, and a reconfigurable SRAM-IMC array. However, there are several aspects that can be further investigated to improve the practical applicability of the proposed designs. The first important direction is complete post-layout validation. Although layout-based area analysis has been performed for selected blocks, a full post-layout extracted evaluation of the proposed sense amplifiers, SRAM bit-cells, SRAM-IMC macro, and reconfigurable array would provide a more accurate estimate of bit-line resistance, interconnect capacitance, coupling effects, routing overhead, and timing skew. This is especially important for SRAM-IMC arrays because the results of this thesis show that array-size scaling mainly increases energy through larger bit-line and peripheral capacitance.

Future work may also focus on larger array organizations beyond the evaluated 32×32 array size. Larger SRAM-IMC arrays would allow the impact of long bit-lines, word-line loading, sense-amplifier sharing, row/column segmentation, local precharge, and hierarchical sensing to be studied in more realistic memory organizations. The present work shows that compute latency

and throughput scale more favorably than compute energy. Therefore, further work should particularly examine energy-aware array techniques such as bit-line segmentation, selective row activation, local compute sub-arrays, reduced-swing evaluation, precharge gating, and peripheral power reduction. These approaches can help reduce average energy per bit while preserving the throughput benefit of SRAM-IMC.

Another important future direction is the development of more robust sensing and timing schemes for compute-mode operation. In conventional SRAM read operation, the sense amplifier distinguishes between two stored states. In SRAM-IMC, the sensing path may have to distinguish bit-line responses corresponding to several operand combinations. Therefore, future work may include offset-aware sensing, adaptive sense-amplifier enable timing, self-timed read/compute control, reference-assisted evaluation, and mismatch-compensated sensing. Such improvements would be useful for maintaining output separability under process-voltage-temperature variations, especially at low supply voltages and in larger arrays.

The proposed SRAM-IMC macro may also be extended beyond NAND/NOR and XOR/XNOR logic. Since these Boolean functions form the basis of many bitwise operations, the same framework can be further developed for bit-serial arithmetic, comparison, population count, addition support, and low-precision multiply–accumulate operations. This would make the proposed design more useful for digital compute-in-memory applications such as binary neural networks, ternary neural networks, search operations, lightweight encryption, and edge-AI preprocessing. In this direction, the main challenge would be to extend the compute functionality without adding excessive sensing, control, or area overhead.

Future research may further include system-level evaluation of the proposed SRAM-IMC framework. A memory-compute circuit can provide useful benefits only when the data mapping, control sequence, and workload requirements are properly matched to the array organization. Therefore, the proposed array can be evaluated using representative bitwise and low-precision workloads to estimate effective throughput, energy per operation, area efficiency, and memory bandwidth reduction. Such analysis would help connect the circuit-level improvement to architecture-level benefits and clarify the application domains where the proposed SRAM-IMC structure is most useful.

The proposed designs can also be studied under additional reliability conditions. The present work includes process-induced variation analysis and Monte Carlo evaluation for different blocks. Future work may include aging effects, supply noise, temperature corners, soft errors, read disturb under repeated compute access, half-select stress in larger arrays, and long-duration retention behavior. Such studies would strengthen the reliability assessment of the proposed SRAM and SRAM-IMC structures, particularly for low-power embedded systems, edge devices, and applications operating under variable environmental conditions.

Finally, fabrication and silicon measurement would be an important step toward practical validation. A test chip containing the proposed sense amplifier, SRAM bit-cell, SRAM-IMC macro, and a small dual-mode SRAM-IMC array would allow the simulated trends to be verified under real process variations. Measured silicon results would provide deeper insight into practical

timing margins, precharge behavior, sense-amplifier offset, compute-mode correctness, leakage, and energy per operation. This would also help identify the additional peripheral and control overhead required for integrating the proposed SRAM-IMC architecture into a larger memory subsystem.

6.3 SOCIAL IMPACT

The proposed work has potential social impact through its contribution to energy-efficient and memory-centric computing. Modern digital systems increasingly process large volumes of data generated by artificial intelligence, communication systems, sensors, mobile devices, and embedded platforms. A large part of the energy and delay in such systems is spent in moving data between memory and logic. SRAM-based in-memory computing directly addresses this issue by allowing selected operations to be performed within or near the memory array. By reducing unnecessary data movement, the proposed design approach can support more energy-efficient digital systems.

The improved sense-amplifier and SRAM bit-cell designs developed in this thesis can also contribute to more reliable low-voltage memory operation. Low-voltage operation is important for reducing dynamic power, but it usually reduces sensing margin and increases sensitivity to device mismatch. The proposed sensing and bit-cell techniques improve read speed, variation tolerance, storage stability, and half-select resilience. These improvements are useful for low-power electronics, where reliable memory operation is required under tight energy constraints.

The proposed SRAM-IMC macro and reconfigurable array can support efficient local processing in edge-computing platforms. Edge devices such as wearable systems, IoT nodes, smart sensors, biomedical monitors, and compact AI accelerators often need to process data close to the source while operating from limited energy resources. By enabling Boolean computation inside the SRAM structure, the proposed architecture can reduce the need for repeated data transfer to external logic. This can improve battery life, reduce access latency, and support faster response in embedded applications.

The work may also support privacy-aware local computation. In many sensor-based systems, raw data are transferred to external processors or cloud platforms for analysis. If simple filtering, comparison, search, or bitwise preprocessing can be performed closer to the memory, the amount of raw data movement can be reduced. Although this thesis does not implement a privacy protocol, the SRAM-IMC direction supports the broader trend of local data processing, which can reduce communication overhead and limit unnecessary exposure of raw data.

The proposed research also contributes to sustainable electronics. Energy-efficient SRAM and SRAM-IMC circuits can reduce the power consumed by memory-bound operations. This is important not only for battery-operated devices but also for large-scale computing systems where repeated memory access contributes significantly to total energy consumption. Even modest reductions in memory-access energy can become meaningful when applied across many devices or many repeated operations. Therefore, the work supports the development of more

energy-aware digital hardware.

The thesis also has educational and research value. It connects four important levels of SRAM design: sense amplifier, bit-cell, SRAM-IMC macro, and reconfigurable SRAM-IMC array. This integrated view can help future researchers understand that SRAM-IMC cannot be optimized by considering only one block. The bit-cell, bit-lines, sense amplifier, compute logic, and array-level parasitics must be considered together. Thus, the work can serve as a useful reference for researchers working on low-power VLSI design, embedded memory, and compute-in-memory architectures.

At the same time, the social impact of this work should be understood realistically. The proposed designs are based on circuit- and array-level simulation and require further post-layout validation, larger-array evaluation, and silicon measurement before practical deployment. Nevertheless, the results provide useful design guidelines for developing faster, more reliable, and more energy-efficient SRAM-IMC systems. In summary, the social relevance of this thesis lies in its contribution toward reducing data-movement overhead, improving low-power memory operation, and supporting future memory-centric computing platforms.

REFERENCES

[1] W. Gul, M. Shams, and D. Al-Khalili, "Sram cell design challenges in modern deep sub-micron technologies: An overview," *Micromachines*, vol. 13, no. 8, p. 1332, 2022.

[2] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, "Modeling of failure probability and statistical design of sram array for yield enhancement in nanoscaled cmos," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, no. 12, pp. 1859–1880, 2005.

[3] S. Mittal, G. Verma, B. Kaushik, and F. A. Khanday, "A survey of sram-based in-memory computing techniques and applications," *Journal of Systems Architecture*, vol. 119, p. 102276, 2021.

[4] Z. Lin, Z. Tong, J. Zhang, F. Wang, T. Xu, Y. Zhao, X. Wu, C. Peng, W. Lu, Q. Zhao, and J. Chen, "A review on sram-based computing in-memory: Circuits, functions, and applications," *Journal of Semiconductors*, vol. 43, no. 3, p. 031401, 2022.

[5] S. Jeloka, N. B. Akesh, D. Sylvester, and D. Blaauw, "A 28 nm configurable memory (tcam/bcam/sram) using push-rule 6t bit cell enabling logic-in-memory," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 4, pp. 1009–1021, 2016.

[6] A. Agrawal, A. Jaiswal, C.-Y. Lee, and K. Roy, "X-sram: Enabling in-memory boolean computations in cmos static random access memories," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 65, no. 12, pp. 4219–4232, 2018.

[7] C.-J. Jhang, C.-X. Xue, J.-M. Hung, F.-C. Chang, and M.-F. Chang, "Challenges and trends of sram-based computing-in-memory for ai edge devices," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 5, pp. 1773–1786, 2021.

[8] A. Garg and T. T.-H. Kim, "Sram array structures for energy efficiency enhancement," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 60, no. 6, pp. 351–355, 2013.

[9] J. Wang, Y. Chen, Z. Wang, Z. Gu, B. Zhang, H. Qiu, K. Zhang, W. Zhao, and Y. Zhang, "A heterogeneous system with computing in memory processing elements to accelerate cnn inference," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 72, no. 7, pp. 3411–3424, 2025.

[10] C. Lammie, H. Benmeziane, W. Simon, E. Ferro, A. Vasilopoulos, J. B"uchel, M. Le Gallo, I. Boybat, and A. Sebastian, "Deep learning software stacks for analogue in-memory computing-based accelerators," *Nature Reviews Electrical Engineering*, vol. 2, no. 9, pp. 621–633, Sep 2025. [Online]. Available: https://doi.org/10.1038/s44287-025-00187-1

[11] T. Yu, T. Liao, M. Zhou, X. Chu, G. Yin, M. Lee, Y. Liu, H. Yang, and X. Li, "Dcirom: A high-density fully-digital compute-in-read-only-memory macro for energy-efficient task-level dnn inference," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 72, no. 12, pp. 7744–7754, 2025.

[12] M. U. K. Sadaf, Z. Chen, S. Subbulakshmi Radhakrishnan, Y. Sun, L. Ding, A. R. Graves, Y. Yang, J. M. Redwing, and S. Das, "Enabling static random-access memory cell scaling with monolithic 3d integration of 2d field-effect transistors," *Nature Communications*, vol. 16, p. 4879, 2025.

[13] S. W. Chung, S. H. Yoon, and J. K. Jeong, "Oxide semiconductor gain cell-embedded memory: materials and integration strategies for next generation on-chip memory," *Communications Engineering*, vol. 5, p. 39, 2026.

[14] T.-Y. J. Chang, Y.-H. Chen, K. V. Reddy, N. Puri, T. Masina, K.-C. Lin, P.-S. Wang, Y. Lin, C.-Y. Lin, Y.-H. Nien, H. Fujiwara, K.-F. Lin, M.-H. Chang, C. W. Wu, R. Lee, Y. Wang, H.-J. Liao, Q. Li, P. W. Wang, and G. Yeap, "A 38.1mb/mm2 sram in a 2nm-cmos-nanosheet technology for high-density and energy-efficient compute," in *2025 IEEE International Solid-State Circuits Conference (ISSCC)*, 2025, pp. 492–494.

[15] Y.-K. Yeh, J.-W. Su, T.-H. Hsu, M. Tseng, J.-C. Tien, K.-C. Chen, C.-Y. Yue, Y.-E. Lin, Y.-J. Hu, L.-J. Hsieh, J.-C. Bai, Y.-S. Kao, T.-H. Lou, H.-H. Hsu, D.-Q. You, S.-S. Sheu, W.-C. Lo, S.-C. Chang, Y.-T. Yang, C.-C. Lo, R.-S. Liu, C.-C. Hsieh, K.-T. Tang, and M.-F. Chang, "A 16nm, 1mb, 1-to-8b-configurable 444.21tops/w fully digital sram compute-in-memory macro for hybrid snn-cnn edge computing," in *2026 IEEE International Solid-State Circuits Conference (ISSCC)*, 2026, pp. 526–528.

[16] C. Dai, J. Zhang, Z. Ren, L. Liu, L. Meng, C. Wang, C. Peng, W. Lu, Z. Lin, and X. Wu, "A t8t-sram computing-in-memory macro for ternary deep neural networks and boolean logic computations," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 34, no. 3, pp. 872–881, 2026.

[17] W. Lu, L. Tan, S. Wang, T. Xue, C. Peng, Z. Lin, and X. Wu, "A 28-nm 9-kb sram computing-in-memory macro with segmented charge sharing for multimode mac operations," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, pp. 1–10, 2026.

[18] T. Chen, Z. Zhu, Y. He, and Q. Li, "An 8-bit precision 10t sram compute-in-memory macro using adc with small area," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 73, no. 2, pp. 113–117, 2026.

[19] H. Wu and Z. Lin, "A 9t sram computation-in-memory architecture with high-precision mac, enhanced bitline voltage margin, and improved frequency performance over conventional architectures," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 34, no. 1, pp. 205–216, 2026.

[20] A. K. Rajput, M. Pattanaik, and G. Kaushal, "An energy-efficient 10t sram in-memory computing macro for artificial intelligence edge processor," *Memories - Materials, Devices, Circuits and Systems*, vol. 5, p. 100076, 2023.

[21] V. Sharma, S. Cosemans, M. Ashouei, J. Huisken, F. Catthoor, and W. Dehaene, "A 4.4 pj/access 80 mhz, 128 kbit variability resilient sram with multi-sized sense amplifier redundancy," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 10, pp. 2416–2430, Oct. 2011.

[22] L. Wen, X. Cheng, K. Zhou, S. Tian, and X. Zeng, "Bit-interleaving-enabled 8t sram with shared data-aware write and reference-based sense amplifier," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 63, no. 7, pp. 643–647, Jul. 2016.

[23] R. Fragasse, R. Tantawy, B. Dupaix, T. Dean, D. DiSabato, M. R. Belz, D. S. Smith, J. J. McCue, and W. Khalil, "Analysis of sram enhancements through sense amplifier capacitive offset correction and replica self-timing," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 66, no. 6, pp. 2037–2050, Jun. 2019.

[24] D. Patel, A. Neale, D. Wright, and M. Sachdev, "Body biased sense amplifier with auto-offset mitigation for low-voltage srams," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 8, pp. 3265–3278, Aug. 2021.

[25] J. Chen, W. Zhao, Y. Wang, and Y. Ha, "Analysis and design of reconfigurable sense amplifier for compute sram with high-speed compute and normal read access," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 68, no. 12, pp. 3503–3507, Dec. 2021.

[26] Y. Zhao, J. Wang, Z. Tong, X. Wu, C. Peng, W. Lu, Q. Zhao, and Z. Lin, "An offset cancellation technique for sram sense amplifier based on relation of the delay and offset," *Microelectronics Journal*, vol. 128, p. 105578, Oct. 2022.

[27] R. Sharma, A. Grover, A. Shroti, and S. Chatterjee, "Up to 45% faster supply boosted voltage sense amplifier (sbvsa) for high-speed srams," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 71, no. 11, pp. 4683–4687, Nov. 2024.

[28] G. Kim, K. Cho, J. Oh, Y. Bae, M. Kim, S. Baeck, T. Song, and S. O. Jung, "Asymmetric voltage latch type and ultra-low swing bitline sense amplifiers for low-power high-density 1r1w 8t sram in 14 nm finfet," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 72, no. 12, pp. 7769–7779, Dec. 2025.

[29] L. G. Heller, D. P. Spampinato, and Y. L. Yao, "High sensitivity charge-transfer sense amplifier," *IEEE Journal of Solid-State Circuits*, vol. 11, no. 5, pp. 596–601, 1976.

[30] T. N. Blalock and R. C. Jaeger, "A high-speed clamped bit-line current-mode sense amplifier," *IEEE Journal of Solid-State Circuits*, vol. 26, no. 4, pp. 542–548, Apr. 1991.

[31] P. Y. Chee, P. C. Liu, and L. Siek, "High-speed hybrid current-mode sense amplifier for cmos srams," *Electronics Letters*, vol. 28, no. 9, pp. 871–873, May 1992.

[32] T. Kobayashi, K. Nogami, T. Shirotori, and Y. Fujimoto, "A current-controlled latch sense amplifier and a static power-saving input buffer for low-power architecture," *IEEE Journal of Solid-State Circuits*, vol. 28, no. 4, pp. 523–527, Apr. 1993.

[33] M. M. Khellah and M. I. Elmasry, "A low-power high-performance current-mode multiport sram," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 9, no. 5, pp. 590–598, Oct. 2001.

[34] B. Wicht, T. Nirschl, and D. Schmitt-Landsiedel, "Yield and speed optimization of a latch-type voltage sense amplifier," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 7, pp. 1148–1158, Jul. 2004.

[35] D. Schinkel, E. Mensink, E. A. M. Klumperink, E. van Tuijl, and B. Nauta, "A double-tail latch-type voltage sense amplifier with 18 ps setup+hold time," in *2007 IEEE International Solid-State Circuits Conference Digest of Technical Papers*, 2007, pp. 314–315.

[36] A. T. Do, Z. H. Kong, and K. S. Yeo, "0.9 v current-mode sense amplifier using concurrent bit- and data-line tracking and sensing techniques," *Electronics Letters*, vol. 43, no. 25, pp. 1421–1422, 2007.

[37] N. Verma and A. P. Chandrakasan, "A 65nm 8t sub-vt sram employing sense-amplifier redundancy," in *2007 IEEE International Solid-State Circuits Conference Digest of Technical Papers*, 2007, pp. 328–329.

[38] A. Kawasumi, Y. Takeyama, O. Hirabayashi, K. Kushida, Y. Fujimura, and T. Yabe, "A low-supply-voltage-operation sram with hci trimmed sense amplifiers," *IEEE Journal of Solid-State Circuits*, vol. 45, no. 11, pp. 2341–2347, 2010.

[39] V. Sharma, S. Cosemans, M. Ashouei, J. Huisken, F. Catthoor, and W. Dehaene, "Ultra low-energy sram design for smart ubiquitous sensors," *IEEE Micro*, vol. 32, no. 5, pp. 10–24, 2012.

[40] D. Patel, A. Neale, D. Wright, and M. Sachdev, "Hybrid latch-type offset tolerant sense amplifier for low-voltage srams," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 66, no. 7, pp. 2519–2532, 2019.

[41] B. S. Reniwal, V. Vijayvargiya, P. Singh, N. K. Yadav, S. K. Vishvakarma, and D. Dwivedi, "An auto-calibrated sense amplifier with offset prediction approach for energy-efficient sram," *Circuits, Systems, and Signal Processing*, vol. 38, pp. 1482–1505, 2019.

[42] D. Dutt, P. Mittal, B. Rawat, and B. Kumar, "Design and performance analysis of high-performance low power voltage mode sense amplifier for static ram," *Advances in Electrical and Electronic Engineering*, vol. 20, no. 3, pp. 285–293, 2022.

[43] S. Shen, H. Xu, Y. Zhou, and W. Yu, "A single-ended offset-canceling sense amplifier enabling wide-voltage operations," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 70, no. 3, pp. 1139–1143, 2023.

[44] L. Chang, D. M. Fried, J. Hergenrother, J. W. Sleight, R. H. Dennard, R. K. Montoye, L. Sekaric, S. J. McNab, A. W. Topol, C. D. Adams, K. W. Guarini, and W. Haensch, "Stable sram cell design for the 32 nm node and beyond," in *Digest of Technical Papers. 2005 Symposium on VLSI Technology*, 2005, pp. 128–129.

[45] K. Takeda, Y. Hagihara, Y. Aimoto, M. Nomura, Y. Nakazawa, T. Ishii, and H. Kobatake, "A read-static-noise-margin-free sram cell for low-vdd and high-speed applications," *IEEE Journal of Solid-State Circuits*, vol. 41, no. 1, pp. 113–121, Jan. 2006.

[46] R. E. Aly and M. A. Bayoumi, "Low-power cache design using 7t sram cell," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 54, no. 4, pp. 318–322, Apr. 2007.

[47] J. P. Kulkarni, K. Kim, and K. Roy, "A 160 mv robust schmitt trigger based subthreshold sram," *IEEE Journal of Solid-State Circuits*, vol. 42, no. 10, pp. 2303–2313, Oct. 2007.

[48] S. Lin, Y.-B. Kim, and F. Lombardi, "A highly-stable nanometer memory for low-power design," in *2008 IEEE International Workshop on Design and Test of Nano Devices, Circuits and Systems*, 2008, pp. 17–20.

[49] J. P. Kulkarni and K. Roy, "Ultralow-voltage process-variation-tolerant schmitt-trigger-based sram design," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 20, no. 2, pp. 319–332, Feb. 2012.

[50] A. S. Rajput, M. Pattanaik, and R. K. Tiwari, "Design and analysis of schmitt trigger based 10t sram in 32 nm technology," in *2017 IEEE International Symposium on Nanoelectronic and Information Systems (iNIS)*, 2017, pp. 234–237.

[51] J. S. Liu, M. B. Clavel, and M. K. Hudait, "An energy-efficient tensile strained ge/ingaas tfet 7t sram cell architecture for ultra-low voltage applications," *IEEE Transactions on Electron Devices*, vol. 64, no. 5, pp. 2193–2200, May 2017.

[52] G. Pasandi and M. Pedram, "Internal write-back and read-before-write schemes to eliminate the disturbance to the half-selected cells in srams," *IET Circuits, Devices & Systems*, vol. 12, no. 4, pp. 460–466, Jul. 2018.

[53] K. Sanapala, S. R, and S. S. Yeo, "Schmitt trigger-based single-ended 7t sram cell for internet of things (iot) applications," *The Journal of Supercomputing*, vol. 74, pp. 4613–4622, Sep. 2018.

[54] R. Giterman, M. Vicentowski, I. Levi, Y. Weizman, O. Keren, and A. Fish, "Leakage power attack-resilient symmetrical 8t sram cell," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 26, no. 10, pp. 2180–2184, Oct. 2018.

[55] S. Pal, S. Bose, W.-H. Ki, and A. Islam, "Characterization of half-select free write assist 9t sram cell," *IEEE Transactions on Electron Devices*, vol. 66, no. 11, pp. 4745–4752, Nov. 2019.

[56] S. Ahmad, S. A. Ahmad, M. Muqeem, N. Alam, and M. Hasan, "Tfet-based robust 7t sram cell for low power application," *IEEE Transactions on Electron Devices*, vol. 66, no. 9, pp. 3834–3840, Sep. 2019.

[57] J. Jiang, Y. Xu, W. Zhu, J. Xiao, and S. Zou, "Quadruple cross-coupled latch-based 10t and 12t sram bit-cell designs for highly reliable terrestrial applications," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 66, no. 3, pp. 967–977, Mar. 2019.

[58] N. Surana and J. Mekie, "Energy efficient single-ended 6-t sram for multimedia applications," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 66, no. 6, pp. 1023–1027, Jun. 2019.

[59] Y. He, J. Zhang, X. Wu, X. Si, S. Zhen, and B. Zhang, "A half-select disturb-free 11t sram cell with built-in write/read-assist scheme for ultralow-voltage operations," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 27, no. 10, pp. 2344–2353, Oct. 2019.

[60] R. N. Asli and S. Taghipour, "Reliable and high performance asymmetric finfet sram cell using back-gate control," *Microelectronics Reliability*, vol. 104, p. 113545, Jan. 2020.

[61] K. Cho, J. Park, T. W. Oh, and S.-O. Jung, "One-sided schmitt-trigger-based 9t sram cell for near-threshold operation," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 67, no. 5, pp. 1551–1561, May 2020.

[62] N. Eslami, B. Ebrahimi, E. Shakouri, and D. Najafi, "A single-ended low leakage and low voltage 10t sram cell with high yield," *Analog Integrated Circuits and Signal Processing*, vol. 105, no. 2, pp. 263–274, Nov. 2020.

[63] E. Shakouri, B. Ebrahimi, N. Eslami, and M. Chahardori, "Single-ended 10t sram cell with high yield and low standby power," *Circuits, Systems, and Signal Processing*, vol. 40, no. 7, pp. 3479–3499, Jul. 2021.

[64] J. Chen, W. Zhao, Y. Wang, Y. Shu, W. Jiang, and Y. Ha, "A reliable 8t sram for high-speed searching and logic-in-memory operations," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 30, no. 6, pp. 769–780, Jun. 2022.

[65] E. Abbasian, "A highly stable low-energy 10t sram for near-threshold operation," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 69, no. 12, pp. 5195–5205, Dec. 2022.

[66] E. Abbasian and S. Sofimowloodi, "A robust multi-bit soft-error immune sram cell for low-power applications," *Analog Integrated Circuits and Signal Processing*, vol. 115, no. 1, pp. 49–66, 2023.

[67] R. G. B. Sangalang, S. Reddy, L. K. S. Tolentino, Y.-W. Shen, O. L. J. A. Jose, and C.-C. Wang, "A 210-mhz 4.23 fj energy/bit 1-kb asymmetrical schmitt-trigger-based sram using 40-nm cmos process," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 70, 2023.

[68] M. Lou, J. Wang, H. Li, Z. Yang, Q. Cheng, J. Li, M. Hashimoto, and L. Lin, "Area-efficient and low-power 8t compute-sram bitcell design for digital compute-in-memory macros in 22-nm cmos," *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2025.

[69] A. Dinesh, Y. Chopra, and P. Mittal, "An 8t sram cell for high-speed sub-threshold operation with dynamic body biasing," *Integration, the VLSI Journal*, vol. 104, p. 102480, Sep. 2025.

[70] A. Srivastav, U. Tiwari, S. K. Mandal, and A. Sachdeva, "A finfet based single bit-line feedback cutting low power 11t sram cell for lpwan applications," *AEU - International Journal of Electronics and Communications*, vol. 197, p. 155821, Jul. 2025.

[71] A. Paul, S. Sharma, and K. Sharma, "Design of a soft error resilient 13t sram architecture for radiation-prone environments in finfet 18 nm technology," *Integration, the VLSI Journal*, vol. 106, p. 102574, 2026.

[72] T. Seki, E. Itoh, C. Furukawa, I. Maeno, T. Ozawa, H. Sano, and N. Suzuki, "A 6-ns 1-mb cmos sram with latched sense amplifier," *IEEE Journal of Solid-State Circuits*, vol. 28, no. 4, pp. 478–483, Apr. 1993.

[73] G. Pasandi and S. M. Fakhraie, "A 256-kb 9t near-threshold sram with 1k cells per bitline and enhanced write and read operations," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 23, no. 11, pp. 2438–2446, 2015.

[74] F. Frustaci, M. Khayatzadeh, D. Blaauw, D. Sylvester, and M. Alioto, "Sram for error-tolerant applications with dynamic energy-quality management in 28 nm cmos," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 5, pp. 1310–1323, 2015.

[75] J. Wang, H. An, Q. Zhang, H. S. Kim, D. Blaauw, and D. Sylvester, "A 40-nm ultra-low leakage voltage-stacked sram for intelligent iot sensors," *IEEE Solid-State Circuits Letters*, vol. 4, pp. 14–17, 2021.

[76] F. Wei, X. Cui, S. Zhang, and X. Zhang, "An 11t sram cell for the dual-direction in-array logic/cam operations," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 71, no. 4, pp. 2329–2333, 2024.

[77] U. Surendranathan, A. S. Vellankanni, A. Milenkovic, U. Guin, and B. Ray, "Ionizing radiation-induced data imprinting effects in sram arrays," *IEEE Transactions on Nuclear Science*, 2025.

[78] J. Zhang, Z. Wang, and N. Verma, "In-memory computation of a machine-learning classifier in a standard 6t sram array," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 4, pp. 915–924, Apr. 2017.

[79] M. Kang, S. K. Gonugondla, A. Patil, and N. R. Shanbhag, "A multi-functional in-memory inference processor using a standard 6t sram array," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 2, pp. 642–655, 2018.

[80] S. Yin, Z. Jiang, J.-S. Seo, and M. Seok, "Xnor-sram: In-memory computing sram macro for binary/ternary deep neural networks," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 6, pp. 1733–1743, 2020.

[81] X. Si, R. Liu, S. Yu, R.-S. Liu, C.-C. Hsieh, K.-T. Tang, Q. Li, M.-F. Chang, J.-J. Chen, Y.-N. Tu, W.-H. Huang, J.-H. Wang, Y.-C. Chiu, W.-C. Wei, S.-Y. Wu, and X. Sun, "A twin-8t sram computation-in-memory unit-macro for multibit cnn-based ai edge processors," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 1, pp. 189–202, Jan. 2020.

[82] J. Mu, H. Kim, and B. Kim, "Sram-based in-memory computing macro featuring voltage-mode accumulator and row-by-row adc for processing neural networks," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 69, no. 6, pp. 2412–2422, Jun. 2022.

[83] V. Damodaran, Z. Liu, J. Meng, J.-S. Seo, and A. Sanyal, "Sram in-memory computing macro with delta-sigma modulator-based variable-resolution activation," *IEEE Solid-State Circuits Letters*, vol. 6, pp. 293–296, 2023.

[84] M. E. Sinangil, B. Erbagci, R. Naous, K. Akarvardar, D. Sun, W.-S. Khwa, H.-J. Liao, Y. Wang, and J. Chang, "A 7-nm compute-in-memory sram macro supporting multi-bit input, weight and output and achieving 351 tops/w and 372.4 gops," *IEEE Journal of Solid-State Circuits*, vol. 56, no. 1, pp. 188–198, Jan. 2021.

[85] Z. Chen, Z. Wen, W. Wan, A. R. Pakala, Y. Zou, W.-C. Wei, Z. Li, Y. Chen, and K. Yang, "Pico-ram: A pvt-insensitive analog compute-in-memory sram macro with in situ multi-bit charge computing and 6t thin-cell-compatible layout," *IEEE Journal of Solid-State Circuits*, vol. 60, no. 1, pp. 308–320, 2025.

[86] Y. Zhan, W.-H. Yu, K.-F. Un, Z. Li, H. Chu, Y. Huan, N. Ma, L.-R. Zheng, R. P. Martins, and P.-I. Mak, "Afp-cim: All-inclusive floating-point with segmented compute-in-memory macro," *IEEE Journal of Solid-State Circuits*, pp. 1–12, 2025.

[87] Q. Dong, S. Jeloka, M. Saligane, Y. Kim, M. Kawaminami, A. Harada, S. Miyoshi, D. Blaauw, and D. Sylvester, "A 0.3v vddmin 4+2t sram for searching and in-memory computing using 55nm ddc technology," in *2017 Symposium on VLSI Circuits*, 2017, pp. C160–C161.

[88] Y.-C. Chiu, Z. Zhang, J.-J. Chen, X. Si, R. Liu, Y.-N. Tu, J.-W. Su, W.-H. Huang, J.-H. Wang, W.-C. Wei, J.-M. Hung, S.-S. Sheu, S.-H. Li, C.-I. Wu, R.-S. Liu, C.-C. Hsieh, K.-T. Tang, and M.-F. Chang, "A 4-kb 1-to-8-bit configurable 6t sram-based computation-in-memory unit-macro for cnn-based ai edge processors," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 10, pp. 2790–2801, Oct. 2020.

[89] C. Eckert, X. Wang, J. Wang, A. Subramaniyan, R. Iyer, D. Sylvester, D. Blaauw, and R. Das, "Neural cache: Bit-serial in-cache acceleration of deep neural networks," in *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*, 2018, pp. 383–396.

[90] T. Abtahi, C. Shea, A. Kulkarni, and T. Mohsenin, "Accelerating convolutional neural network with fft on embedded hardware," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 26, no. 9, pp. 1737–1749, Sep. 2018.

[91] S. K. Gonugondla, M. Kang, and N. R. Shanbhag, "A variation-tolerant in-memory machine learning classifier via on-chip training," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 11, pp. 3163–3173, Nov. 2018.

[92] M. Ali, A. Jaiswal, S. Kodge, A. Agrawal, I. Chakraborty, and K. Roy, "Imac: In-memory multi-bit multiplication and accumulation in 6t sram array," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 67, no. 8, pp. 2521–2531, Aug. 2020.

[93] A. Biswas and A. P. Chandrakasan, "Conv-sram: An energy-efficient sram with in-memory dot-product computation for low-power convolutional neural networks," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 1, pp. 217–230, Jan. 2019.

[94] J. Kim, J. Koo, T. Kim, Y. Kim, H. Kim, S. Yoo, and J. J. Kim, "Area-efficient and variation-tolerant in-memory bnn computing using 6t sram array," in *2019 Symposium on VLSI Circuits, Digest of Technical Papers*, 2019, pp. C118–C119.

[95] M. Kang and collaborators, "Sram-based cim array for neural acceleration," 2020, source metadata incomplete in current materials; verify against final manuscript before thesis submission.

[96] K. Lee, J. Jeong, S. Cheon, W. Choi, and J. Park, "Bit parallel 6t sram in-memory computing with reconfigurable bit-precision," in *2020 57th ACM/IEEE Design Automation Conference (DAC)*, 2020, pp. 1–6.

[97] H. W. Jiang, X. C. Peng, S. S. Huang, and S. M. Yu, "Cimat: A compute-in-memory architecture for on-chip training based on transpose sram arrays," *IEEE Transactions on Computers*, vol. 69, no. 7, pp. 944–954, Jul. 2020.

[98] J.-W. Su, X. Si, Y.-C. Chou, T.-W. Chang, W. Huang, Y. Tu, R. Liu, P. Lu, T. Liu, J. Wang, Z. Zhang, H. Jiang, S. Huang, C. Lo, R. Liu, C. Hsieh, K. T. S. Tang, S. S. Sheu, S. Li, H. Lee, S. Chang, S. Yu, and M. F. Chang, "A 28nm 64kb inference-training two-way transpose multibit 6t sram compute-in-memory macro for ai edge chips," in *2020 IEEE International Solid-State Circuits Conference (ISSCC)*, Feb. 2020, pp. 240–242.

[99] J.-W. Su, Y.-C. Chou, R. Liu, T.-W. Liu, P.-J. Lu, P.-C. Wu, Y.-L. Chung, L.-Y. Hung, J.-S. Ren, T. Pan, S.-H. Li, S.-C. Chang, S.-S. Sheu, W.-C. Lo, C.-I. Wu, X. Si, C.-C. Lo, R.-S. Liu, C.-C. Hsieh, K.-T. Tang, and M.-F. Chang, "16.3 a 28nm 384kb 6t-sram computation-in-memory macro with 8b precision for ai edge chips," in *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, 2021, pp. 250–252.

[100] K. Lee, J. Kim, and J. Park, "Low-cost 7t-sram compute-in-memory design based on bit-line charge-sharing based analog-to-digital conversion," in *2022 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 2022.

[101] Y. Zhang, M. Wang, Y. Mai, and Z. Yu, "Tensorcache: Reconstructing memory architecture with sram-based in-cache computing for efficient tensor computations in gpgpus," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 31, no. 12, pp. 2030–2043, Dec. 2023.

[102] K. Xiao, X. Cui, X. Qiao, J. Song, H. Luo, X. Wang, and Y. Wang, "A 28nm 32kb sram computing-in-memory macro with hierarchical capacity attenuator and input sparsity-optimized adc for 4b mac operation," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 70, no. 5, pp. 1816–1820, May 2023.

[103] X. Li and collaborators, "A 9t-sram based computing-in-memory with redundant unit and digital operation for boolean logic and mac," *Microelectronics Journal*, vol. 145, p. 106124, 2024.

[104] K. Lee, J. Kim, and J. Park, "A 28-nm 50.1-tops/w p-8t sram compute-in-memory macro design with bl charge-sharing-based in-sram dac/adc operations," *IEEE Journal of Solid-State Circuits*, vol. 59, no. 6, pp. 1926–1937, Jun. 2024.

[105] J. Park, J. Rhe, C. Hwang, J. So, and J. H. Ko, "Input/mapping precision controllable digital cim with adaptive adder tree architecture for flexible dnn inference," *Journal of Systems Architecture*, vol. 159, p. 103327, 2025.

[106] J. Wang, X. Wang, C. Eckert, A. Subramaniyan, R. Das, D. Blaauw, and D. Sylvester, "A 28-nm compute sram with bit-serial logic/arithmetic operations for programmable in-memory vector computing," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 1, pp. 76–86, Jan. 2020.

[107] H. Zhu and V. Kursun, "A comprehensive comparison of data stability enhancement techniques with novel nanoscale sram cells under parameter fluctuations," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 61, no. 5, pp. 1473–1484, 2014.

[108] B. Rawat and P. Mittal, "A reliable and temperature variation tolerant 7t sram cell with single bitline configuration for low voltage application," *Circuits, Systems, and Signal Processing*, vol. 41, no. 5, pp. 2779–2801, 2022.

[109] ——, "A comprehensive analysis of different 7T SRAM topologies to design a 1R1W bit interleaving enabled and half select free cell for 32 nm technology node," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 478, no. 2259, p. 20210745, 2022.

[110] ——, "Single bit line accessed high-performance ultra-low voltage operating 7t static random access memory cell with improved read stability," *International Journal of Circuit Theory and Applications*, vol. 49, no. 5, pp. 1435–1449, 2021.

[111] ——, "A reliable and temperature variation tolerant 7t sram cell with single bitline configuration for low voltage application," *Circuits, Systems, and Signal Processing*, vol. 41, no. 5, pp. 2779–2801, 2022.

[112] R. Lorenzo and R. Paily, "Half-selection disturbance free 8t low leakage sram cell," *International Journal of Circuit Theory and Applications*, vol. 50, no. 5, pp. 1557–1575, 2022.

[113] E. Abbasian and M. Gholipour, "A low-leakage single-bitline 9t sram cell with read-disturbance removal and high writability for low-power biomedical applications," *International Journal of Circuit Theory and Applications*, vol. 50, no. 5, pp. 1537–1556, 2022.

[114] P. Mittal, B. Rawat, and N. Kumar, "Tetra-variate scrutiny of diverse multiplexer techniques for designing a barrel shifter for low power digital circuits," *Microprocessors and Microsystems*, vol. 90, p. 104491, 2022.

[115] H. Reyserhove and W. Dehaene, "A differential transmission gate design flow for minimum energy sub-10-pj/cycle arm cortex-m0 mcus," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 7, pp. 1904–1914, 2017.

[116] B. Rawat and P. Mittal, "A switching nmos based single ended sense amplifier for high density sram applications," *ACM Trans. Des. Autom. Electron. Syst.*, dec 2022, just Accepted.

[117] S. Shen, H. Xu, Y. Zhou, and W. Yu, "A single-ended offset-canceling sense amplifier enabling wide-voltage operations," *IEEE Transactions on Circuits and Systems II: Express Briefs*, p. 1, 2022.

[118] D. Dutt, P. Mittal, B. Rawat, and B. Kumar, "Design and performance analysis of high-performance low power voltage mode sense amplifier for static ram," *Advances in Electrical and Electronic Engineering*, vol. 20, no. 3, pp. 285–293, 2022.

[119] B. Wicht, T. Nirschl, and D. Schmitt-Landsiedel, "Yield and speed optimization of a latch-type voltage sense amplifier," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 7, pp. 1148–1158, 2004.

[120] T. Kobayashi, K. Nogami, T. Shirotori, and Y. Fujimoto, "A current-controlled latch sense amplifier and a static power-saving input buffer for low-power architecture," *IEICE transactions on electronics*, vol. 76, no. 5, pp. 863–867, 1993.

[121] D. Schinkel, E. Mensink, E. Klumperink, E. Van Tuijl, and B. Nauta, "A double-tail latch-type voltage sense amplifier with 18ps setup+ hold time," in *2007 IEEE international solid-state circuits conference. Digest of technical papers*. IEEE, 2007, pp. 314–605.

[122] T. N. Blalock and R. C. Jaeger, "A high-speed clamped bit-line current-mode sense amplifier," *IEEE Journal of Solid-State Circuits*, vol. 26, no. 4, pp. 542–548, 1991.

[123] P. Chee, P. Liu, and L. Siek, "High-speed hybrid current-mode sense amplifier for cmos srams," *Electronics Letters*, vol. 9, no. 28, pp. 871–873, 1992.

[124] J.-S. Wang and H.-Y. Lee, "A new current-mode sense amplifier for low-voltage low-power sram," in *Proceedings Eleventh Annual IEEE International ASIC Conference (Cat. No. 98TH8372)*. IEEE, 1998, pp. 163–167.

[125] M. M. Khellah and M. I. Elmasry, "A low-power high-performance current-mode multiport sram," *IEEE transactions on very large scale integration (VLSI) systems*, vol. 9, no. 5, pp. 590–598, 2001.

[126] A. Do, Z. Kong, and K. Yeo, "0.9 v current-mode sense amplifier using concurrent bit-and data-line tracking and sensing techniques," *Electronics Letters*, vol. 43, no. 25, pp. 1421–1422, 2007.

[127] E. Seevinck, P. J. van Beers, and H. Ontrop, "Current-mode techniques for high-speed vlsi circuits with application to current sense amplifier for cmos sram's," *IEEE Journal of Solid-State Circuits*, vol. 26, no. 4, pp. 525–536, 1991.

[128] J. S. Shah, D. Nairn, and M. Sachdev, "An energy-efficient offset-cancelling sense amplifier," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 60, no. 8, pp. 477–481, 2013.

[129] A. J. Bhavnagarwala, X. Tang, and J. D. Meindl, "The impact of intrinsic device fluctuations on cmos sram cell stability," *IEEE Journal of Solid-State Circuits*, vol. 36, 2001.

[130] D. Patel, A. Neale, D. Wright, and M. Sachdev, "Body biased sense amplifier with auto-offset mitigation for low-voltage srams," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 8, pp. 3265–3278, 2021.

[131] B. S. Reniwal, V. Vijayvargiya, P. Singh, N. K. Yadav, S. K. Vishvakarma, and D. Dwivedi, "An auto-calibrated sense amplifier with offset prediction approach for energy-efficient sram," *Circuits, Systems, and Signal Processing*, vol. 38, 2019.

[132] R. Fragasse, R. Tantawy, B. Dupaix, T. Dean, D. Disabato, M. R. Belz, D. Smith, J. McCue, and W. Khalil, "Analysis of sram enhancements through sense amplifier capacitive offset correction and replica self-timing," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 66, 2019.

[133] Y. Zhao, J. Wang, Z. Tong, X. Wu, C. Peng, W. Lu, Q. Zhao, and Z. Lin, "An offset cancellation technique for sram sense amplifier based on relation of the delay and offset," *Microelectronics Journal*, vol. 128, p. 105578, 10 2022.

[134] N. Verma and A. P. Chandrakasan, "A 65nm 8t sub-vt sram employing sense-amplifier redundancy," in *2007 IEEE International Solid-State Circuits Conference. Digest of Technical Papers*. IEEE, 2007, pp. 328–606.

[135] V. Sharma, S. Cosemans, M. Ashouei, J. Huisken, F. Catthoor, and W. Dehaene, "A 4.4 pj/access 80 mhz, 128 kbit variability resilient sram with multi-sized sense amplifier redundancy," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 10, pp. 2416–2430, 2011.

[136] J. Chen, W. Zhao, Y. Wang, and Y. Ha, "Analysis and optimization strategies toward reliable and high-speed 6t compute sram," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 4, pp. 1520–1531, 2021.

[137] C. Hernández, A. Roca, J. Flich, F. Silla, and J. Duato, "Characterizing the impact of process variation on 45 nm noc-based cmps," *Journal of Parallel and Distributed Computing*, vol. 71, no. 5, pp. 651–663, 2011.

[138] T. Kobayashi, K. Nogami, T. Shirotori, and Y. Fujimoto, "A current-controlled latch sense amplifier and a static power-saving input buffer for low-power architecture," *IEICE Transactions on Electronics*, vol. 76, no. 5, pp. 863–867, 1993.

[139] B. Wicht, T. Nirschl, and D. Schmitt-Landsiedel, "Yield and speed optimization of a latch-type voltage sense amplifier," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 7, pp. 1148–1158, 2004.

[140] D. Schinkel, E. Mensink, E. Klumperink, E. Van Tuijl, and B. Nauta, "A double-tail latch-type voltage sense amplifier with 18ps setup+ hold time," in *2007 IEEE International Solid-State Circuits Conference. Digest of Technical Papers*, 2007, pp. 314–605.

[141] T. N. Blalock and R. C. Jaeger, "A high-speed clamped bit-line current-mode sense amplifier," *IEEE Journal of Solid-State Circuits*, vol. 26, no. 4, pp. 542–548, 1991.

[142] P. Y. Chee, P. C. Liu, and L. Siek, "High-speed hybrid current-mode sense amplifier for cmos srams," *Electronics Letters*, vol. 28, no. 9, pp. 871–873, 1992.

[143] M. M. Khellah and M. I. Elmasry, "A low-power high-performance current-mode multiport sram," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 9, no. 5, pp. 590–598, 2001.

[144] A. T. Do, Z. H. Kong, and K. S. Yeo, "0.9 v current-mode sense amplifier using concurrent bit-and data-line tracking and sensing techniques," *Electronics Letters*, vol. 43, no. 25, pp. 1421–1422, 2007.

[145] N. I. Arizona State University and M. Group, "Predictive Technology Models," 2008. [Online]. Available: http://ptm.asu.edu/latest.html

[146] J. Zhu, N. Bai, and J. Wu, "A review of sense amplifiers for static random access memory," *IETE Technical Review*, vol. 30, no. 1, pp. 72–81, 2013.

[147] M. Strobel and M. Radetzki, "Power-mode-aware memory subsystem optimization for low-power system-on-chip design," *ACM Trans. Embed. Comput. Syst.*, vol. 18, no. 5, Oct. 2019. [Online]. Available: https://doi.org/10.1145/3356583

[148] S. Pal, S. Bose, W.-H. Ki, and A. Islam, "Reliable write assist low power sram cell for wireless sensor network applications," *IET Circuits, Devices & Systems*, vol. 14, no. 2, pp. 137–147, 2020. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-cds.2019.0050

[149] H. Jayakumar, A. Raha, J. R. Stevens, and V. Raghunathan, "Energy-aware memory mapping for hybrid fram-sram mcus in intermittently-powered iot devices," *ACM Trans. Embed. Comput. Syst.*, vol. 16, no. 3, Apr. 2017. [Online]. Available: https://doi.org/10.1145/2983628

[150] J. Kwong and A. Chandrakasan, "An energy-efficient biomedical signal processing platform," *Solid-State Circuits, IEEE Journal of*, vol. 46, pp. 1742 – 1753, 08 2011.

[151] H. Jayakumar, A. Raha, and V. Raghunathan, "Sleep-mode voltage scaling: Enabling sram data retention at ultra-low power in embedded microcontrollers," *ACM Trans. Embed. Comput. Syst.*, vol. 16, no. 1, Nov. 2016. [Online]. Available: https://doi.org/10.1145/2950054

[152] C.-F. Hsu, H.-S. Zheng, Y.-Y. Liu, and T. T. Yeh, "Streamnet++: Memory-efficient streaming tinyml model compilation on microcontrollers," *ACM Trans. Embed. Comput. Syst.*, vol. 24, no. 2, Jan. 2025. [Online]. Available: https://doi.org/10.1145/3706107

[153] R. Yao, H. Lv, Y. Zhang, X. Chen, Y. Zhang, X. Liu, and G. Bai, "A high-reliability 12t sram radiation-hardened cell for aerospace applications," *Micromachines*, vol. 14, no. 7, 2023. [Online]. Available: https://www.mdpi.com/2072-666X/14/7/1305

[154] B. Rawat and P. Mittal, "A comprehensive analysis of different 7t sram topologies to design a 1r1 w bit interleaving enabled and half select free cell for 32 nm technology node," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 478, no. 2259, p. 20210745, 2022. [Online]. Available: https://royalsocietypublishing.org/doi/abs/10.1098/rspa.2021.0745

[155] S. Gupta, K. Gupta, B. H. Calhoun, and N. Pandey, "Low-power near-threshold 10t sram bit cells with enhanced data-independent read port leakage for array augmentation in 32-nm cmos," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 66, no. 3, pp. 978–988, 2019.

[156] J. Kulkarni, K. Kim, and K. Roy, "A 160 mv robust schmitt trigger based subthreshold sram," *Solid-State Circuits, IEEE Journal of*, vol. 42, pp. 2303 – 2313, 11 2007.

[157] E. Abbasian, M. Gholipour, and S. Birla, "A single-bitline 9t sram for low-power near-threshold operation in finfet technology," *Arabian Journal for Science and Engineering*, vol. 47, no. 11, pp. 14 543–14 559, Nov 2022. [Online]. Available: https://doi.org/10.1007/s13369-022-06821-6

[158] B. Rawat and P. Mittal, "A low power single bit-line configuration dependent 7t sram bit cell with process-variation-tolerant enhanced read performance," *Analog Integrated Circuits and Signal Processing*, vol. 115, no. 1, pp. 77–92, Apr 2023.

[159] A. Sachdeva and V. Tomar, "Design of multi-cell upset immune single-end sram for low power applications," *AEU - International Journal of Electronics and Communications*, vol. 128, p. 153516, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1434841120322044

[160] V. Singhal, B. Rawat, P. Mittal, and B. Kumar, "A single ended, single port configuration based 9 t sram cell for stability enhancement," *Physica Scripta*, vol. 98, no. 11, p. 115035, oct 2023. [Online]. Available: https://dx.doi.org/10.1088/1402-4896/ad00e1

[161] E. Abbasian and M. Gholipour, "Design of a schmitt-trigger-based 7t sram cell for variation resilient low-energy consumption and reliable internet of things applications," *AEU - International Journal of Electronics and Communications*, vol. 138, p. 153899, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S143484112100296X

[162] E. Abbasian, E. Mani, M. Gholipour, M. Karamimanesh, M. Sahid, and A. Zaidi, "A schmitt-trigger-based low-voltage 11 t sram cell for low-leakage in 7-nm finfet technology," *Circuits, Systems, and Signal Processing*, vol. 41, no. 6, pp. 3081–3105, Jun 2022. [Online]. Available: https://doi.org/10.1007/s00034-021-01950-z

[163] E. Abbasian and M. Gholipour, "Single-ended half-select disturb-free 11t static random access memory cell for reliable and low power applications," *International Journal of Circuit Theory and Applications*, vol. 49, no. 4, pp. 970–989, 2021.

[164] Y.-W. Chiu, Y.-H. Hu, M.-H. Tu, J.-K. Zhao, Y.-H. Chu, S.-J. Jou, and C.-T. Chuang, "40 nm bit-interleaving 12t subthreshold sram with data-aware write-assist," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 61, pp. 2578–2585, 09 2014.

[165] J. Singh, S. P. Mohanty, and D. K. Pradhan, *Robust SRAM designs and analysis*. Springer Science & Business Media, 2012.

[166] A. Islam and M. Hasan, "Leakage characterization of 10t sram cell," *IEEE transactions on electron devices*, vol. 59, no. 3, pp. 631–638, 2012.

[167] M. Ali, A. Jaiswal, S. Kodge, A. Agrawal, I. Chakraborty, and K. Roy, "Imac: In-memory multi-bit multiplication and accumulation in 6t sram array," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 67, no. 8, pp. 2521–2531, 2020.

[168] C. Kushwah, S. K. Vishvakarma, and D. Dwivedi, "Single-ended boost-less (se-bl) 7t process tolerant sram design in sub-threshold regime for ultra-low-power applications," *Circuits, Systems, and Signal Processing*, vol. 35, no. 2, pp. 385–407, 2016.

[169] P. Singh and S. K. Vishvakarma, "Ultra-low power high stability 8t sram for application in object tracking system," *IEEE Access*, vol. 6, pp. 2279–2290, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:3296952

[170] C. Kushwah and S. K. Vishvakarma, "A single-ended with dynamic feedback control 8t subthreshold sram cell," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 24, no. 1, pp. 373–377, 2015.

[171] L. Chang, D. Fried, J. Hergenrother, J. Sleight, R. Dennard, R. Montoye, L. Sekaric, S. McNab, A. Topol, C. Adams, K. Guarini, and W. Haensch, "Stable sram cell design for the 32 nm node and beyond," in *Digest of Technical Papers. 2005 Symposium on VLSI Technology, 2005.*, 2005, pp. 128–129.

[172] V. Gupta and M. Anis, "Statistical design of the 6t sram bit cell," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 57, no. 1, pp. 93–104, 2009.

[173] C. Roy and A. Islam, "Characterization of single-ended 9t sram cell," *Microsystem Technologies*, vol. 26, pp. 1591–1604, 2020.

[174] N. Eslami, B. Ebrahimi, E. Shakouri, and D. Najafi, "A single-ended low leakage and low voltage 10t sram cell with high yield," *Analog Integrated Circuits and Signal Processing*, vol. 105, pp. 263–274, 2020.

[175] R. Lorenzo and R. Pailly, "Single bit-line 11t sram cell for low power and improved stability," *IET Computers & Digital Techniques*, vol. 14, no. 3, pp. 114–121, 2020.

[176] Q. Zhao, C. Peng, J. Chen, Z. Lin, and X. Wu, "Novel write-enhanced and highly reliable rhpd-12t sram cells for space applications," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 28, no. 3, pp. 848–852, 2020.

[177] M. AlSelek, J. M. Alcaraz-Calero, and Q. Wang, "Dynamic ai-iot: Enabling updatable ai models in ultralow-power 5g iot devices," *IEEE Internet of Things Journal*, vol. 11, no. 8, pp. 14 192–14 205, 2024.

[178] H.-S. P. Wong and S. Salahuddin, "Memory leads the way to better computing," *Nature Nanotechnology*, vol. 10, no. 3, pp. 191–194, Mar 2015. [Online]. Available: https://doi.org/10.1038/nnano.2015.29

[179] J. Zeng, N. Xu, Y. Chen, C. Huang, Z. Li, and L. Fang, "Aimcu-meso: An in-memory computing unit constructed by meso device," *ACM Trans. Des. Autom. Electron. Syst.*, vol. 28, no. 1, dec 2022. [Online]. Available: https://doi.org/10.1145/3539575

[180] A. Agrawal, A. Jaiswal, C. Lee, and K. Roy, "X-sram: Enabling in-memory boolean computations in cmos static random access memories," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 65, no. 12, pp. 4219–4232, 2018.

[181] J. Chen, W. Zhao, Y. Wang, Y. Shu, W. Jiang, and Y. Ha, "A reliable 8t sram for high-speed searching and logic-in-memory operations," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 30, no. 6, pp. 769–780, 2022.

[182] Z. Lin, S. Zhang, Q. Jin, J. Xia, Y. Liu, K. Yu, J. Zheng, X. Xu, X. Fan, K. Li, Z. Tong, X. Wu, W. Lu, C. Peng, and Q. Zhao, "A fully digital sram-based four-layer in-memory computing unit achieving multiplication operations and results store," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 31, no. 6, pp. 776–788, 2023.

[183] B. Rawat and P. Mittal, "A reconfigurable 7t sram bit cell for high speed, power saving and low voltage application," *ACM Trans. Des. Autom. Electron. Syst.*, vol. 28, no. 6, oct 2023. [Online]. Available: https://doi.org/10.1145/3616872

[184] W. He, S. Yin, Y. Kim, X. Sun, J.-J. Kim, S. Yu, and J.-S. Seo, "2-bit-per-cell rram-based in-memory computing for area-/energy-efficient deep learning," *IEEE Solid-State Circuits Letters*, vol. 3, pp. 194–197, 2020.

[185] W. Wan, R. Kubendran, C. Schaefer, S. B. Eryilmaz, W. Zhang, D. Wu, S. Deiss, P. Raina, H. Qian, B. Gao, S. Joshi, H. Wu, H.-S. P. Wong, and G. Cauwenberghs, "A compute-in-memory chip based on resistive random-access memory," *Nature*, vol. 608, no. 7923, pp. 504–512, Aug 2022. [Online]. Available: https://doi.org/10.1038/s41586-022-04992-8

[186] C. Yu, T. Yoo, H. Kim, T. T.-H. Kim, K. C. T. Chuan, and B. Kim, "A logic-compatible edram compute-in-memory with embedded adcs for processing neural networks," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 2, pp. 667–679, 2021.

[187] F. Zhang, S. Angizi, J. Sun, W. Zhang, and D. Fan, "Aligner-d: Leveraging in-dram computing to accelerate dna short read alignment," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 13, no. 1, pp. 332–343, 2023.

[188] J. Wu, Y. Wang, P. Wang, Y. Wang, and W. Zhao, "A stt-assisted sot mram-based in-memory booth multiplier for neural network applications," *IEEE Transactions on Nanotechnology*, vol. 23, pp. 29–34, 2024.

[189] B. Wu, K. Liu, T. Yu, H. Zhu, K. Chen, C. Yan, E. Deng, and W. Liu, "High-performance stt-mram-based computing-in-memory scheme utilizing data read feature," *IEEE Transactions on Nanotechnology*, vol. 22, pp. 817–826, 2023.

[190] Z. Xuan, C. Liu, Y. Zhang, Y. Li, and Y. Kang, "A brain-inspired adc-free sram-based in-memory computing macro with high-precision mac for ai application," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 70, no. 4, pp. 1276–1280, 2023.

[191] L. Han, P. Huang, Y. Wang, Z. Zhou, Y. Zhang, X. Liu, and J. Kang, "Efficient discrete temporal coding spike-driven in-memory computing macro for deep neural network based on nonvolatile memory," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 69, no. 11, pp. 4487–4498, 2022.

[192] Z. Jiang, S. Yin, J.-S. Seo, and M. Seok, "C3sram: An in-memory-computing sram macro based on robust capacitive coupling computing mechanism," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 7, pp. 1888–1897, 2020.

[193] S. Yin, Z. Jiang, J.-S. Seo, and M. Seok, "Xnor-sram: In-memory computing sram macro for binary/ternary deep neural networks," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 6, pp. 1733–1743, 2020.

[194] J. P. Kulkarni, A. Goel, P. Ndai, and K. Roy, "A read-disturb-free, differential sensing 1r/1w port, 8t bitcell array," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 19, no. 9, pp. 1727–1730, 2011.

[195] S. Jeloka, N. B. Akesh, D. Sylvester, and D. Blaauw, "A 28 nm configurable memory (tcam/bcam/sram) using push-rule 6t bit cell enabling logic-in-memory," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 4, pp. 1009–1021, 2016.

[196] S. Pal and A. Islam, "Variation tolerant differential 8t sram cell for ultralow power applications," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 35, no. 4, pp. 549–558, 2016.

[197] C. Dai, Z. Ren, L. Guan, H. Liu, M. Gao, W. Lu, Z. Pang, C. Peng, and X. Wu, "A 9t-sram in-memory computing macro for boolean logic and multiply-and-accumulate operations," *Microelectronics Journal*, vol. 144, p. 106087, 2024.

[198] Y. Zhao, Y. Liu, J. Zheng, Z. Tong, X. Wang, R. Yu, X. Wu, Y. Zhou, C. Peng, W. Lu, Q. Zhao, and Z. Lin, "Configurable in-memory computing architecture based on dual-port sram," *Microelectronics Journal*, vol. 147, p. 106163, 2024.

[199] Z. Lin, R. Yu, D. Huo, Q. Zhu, M. Long, Y. Qin, Y. Liu, L. Chen, S. Wang, T. Wang, Y. Xing, Z. Wen, Y. Liu, X. Li, C. Dai, Q. Zhao, C. Peng, and X. Wu, "A 28-nm 9t sram-based cim macro with capacitance weighting module and redundant array-assisted adc," *Microelectronics Journal*, vol. 153, p. 106397, 2024.

[200] C. Dai, J. Zhang, R. Wang, J. Chen, W. Hu, L. Hao, W. Lu, Z. Lin, C. Peng, and X. Wu, "A pvt-insensitive 7t sram cim macro for multibit multiplication with dynamic matching quantization circuits," *Microelectronics Journal*, vol. 161, p. 106703, 2025.

[201] H. Kim, J. Mu, C. Yu, T. T. H. Kim, and B. Kim, "A 1–16b reconfigurable 80kb 7t sram-based digital near-memory computing macro for processing neural networks," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 70, no. 4, pp. 1580–1590, 2023.

LIST OF PUBLICATIONS

SCI/SCIE-Indexed International Journal Papers

1. Ayush Dahiya, Poornima Mittal, and Rajesh Rohilla, "Variation-Tolerant Sense Amplifier Using Decoupling Transistors for Enhanced SRAM Read Performance," *Circuits, Systems, and Signal Processing*, vol. 42, no. 10, pp. 5799–5810, May 2023, doi: 10.1007/s00034-023-02397-0. **[IF: 2.0, Objective – 1]**

2. Ayush Dahiya, Poornima Mittal, and Rajesh Rohilla, "Modified Decoupled Sense Amplifier with Improved Sensing Speed for Low-Voltage Differential SRAM," *ACM Transactions on Design Automation of Electronic Systems*, vol. 28, no. 6, Article 94, pp. 1–15, November 2023, doi: 10.1145/3611672. **[IF: 2.2, Objective – 1]**

3. Ayush Dahiya, Poornima Mittal, and Rajesh Rohilla, "Realizing In-Memory Computing Using Reliable Differential 8T SRAM for Improved Latency," *ACM Transactions on Design Automation of Electronic Systems*, vol. 29, no. 6, Article 106, pp. 1–15, November 2024, doi: 10.1145/3696666. **[IF: 2.2, Objective – 3]**

4. Ayush Dahiya, Poornima Mittal, and Rajesh Rohilla, "A Low-Power Half-Select Free 8T SRAM Cell with Process-Induced Variation Resistance for Voltage Scaling at 32 nm Technology Node," *Integration*, vol. 106, Article 102586, November 2025, doi: 10.1016/j.vlsi.2025.102586. **[IF: 2.5, Objective – 2]**

Scopus-Indexed International Conference Papers

1. Ayush Dahiya, Poornima Mittal, and Rajesh Rohilla, "Comparative Analysis of Current Sense Amplifier Architectures for SRAM at 45 nm Technology Node," *International Conference on Advances in Data-Driven Computing and Intelligent Systems (ADCIS 2022)*, organized by BITS Pilani, Goa, India, September 23–25, 2022. **[Objective – 1]**

2. Ayush Dahiya, Poornima Mittal, and Rajesh Rohilla, "Design Trade-Offs in 8T SRAM Architectures: Differential and Single-Ended Approaches," *6th International Conference on Data Science and Applications (ICDSA 2025)*, organized by MNIT Jaipur, Rajasthan, India, July 16–18, 2025. **[Objective – 2]**

Circuits, Systems, and Signal Processing (2023) 42:5799–5810
https://doi.org/10.1007/s00034-023-02397-0

Variation-Tolerant Sense Amplifier Using Decoupling Transistors for Enhanced SRAM Read Performance

Ayush Dahiya[1] · **Poornima Mittal**[1] · **Rajesh Rohilla**[1]

Received: 19 November 2022 / Revised: 30 April 2023 / Accepted: 2 May 2023 /
Published online: 23 May 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract

A decoupled inverter pair sense amplifier (DIPSA) is proposed for improved sensing in differential SRAM. The DIPSA is designed at 45-nm technology node with a decreased sensing delay of 31.44% and 28.44% over the conventional latch-type sense amplifier (CLSA) and current-latched SA with an NMOS footswitch (CSANF) at 1.1 V, respectively. The performance of DIPSA is validated using delay and power consumption comparison with prior SA designs across a set of supply voltages. Monte Carlo simulation for 3σ deviation yields a reliable design with no read failures and mean delay of 488.81 ps across the 1k data points with a deviation of 54.87 ps. The threshold voltage variation contribution to process variation is considered for these calculations. The sizing of the decoupling transistors is discussed and its impact on the delay and power of DIPSA shown. The performance of existing voltage and current sense amplifiers is compared to DIPSA.

Keywords Low-voltage SRAM · Sensing delay · Sense amplifier · SRAM sensing · Transient analysis · Variation-tolerant circuits

1 Introduction

The SRAM forms a dominant part of most modern computational systems-on-chip (SoCs). The SRAM component covers larger area compared to the logic circuits in

✉ Poornima Mittal
poornimamittal@dtu.ac.in

Ayush Dahiya
ayushdahiya010@gmail.com

Rajesh Rohilla
rajesh@dce.ac.in

[1] Department of Electronics and Communication Engineering, Delhi Technological University, Delhi 110042, India

Modified Decoupled Sense Amplifier with Improved Sensing Speed for Low-Voltage Differential SRAM

AYUSH DAHIYA, POORNIMA MITTAL, and RAJESH ROHILLA, Delhi Technological University, India

A modified decoupled sense amplifier (MDSA) and modified decoupled sense amplifier with NMOS footswitch is proposed for improved sensing in differential SRAM for low-voltage operation at the 22-nm technology node. The MDSA and MDSANF both offer notable improvements to read delay over conventional voltage and current sense amplifiers. At an operating voltage of 0.8 V, the MDSA exhibited a reduced delay of 28.6%, 41.79%, 37.74%, and 30.94% compared to modified clamped sense amplifier (MCSA), double tail sense amplifier (DTSA), modified hybrid sense amplifier (MHSA), and conventional latch-type sense amplifier (LSA), respectively. Similarly, the MDSANF demonstrated a delay reduction of 26.13%, 39.78%, 35.58%, and 28.55% over MCSA, DTSA, MHSA, and LSA, respectively. To validate the performance, the MDSA and MDSANF are evaluated using the variation in delay and power consumption across various supply voltages, process corners, input differential bit line voltage (ΔV_{BL}), bit line capacitance (C_{BL}), and the sizing of decoupling transistors. Monte Carlo simulations were conducted to analyse the impact of voltage threshold variations on transistor mismatch which leads to an increased occurrence of read failures and a decline in SRAM yield. The performance analysis of various voltage and current sense amplifiers is presented along with MDSA and MDSANF. Area consideration for selection of sensing scheme is important and as such layout of MDSA and MDSANF was performed conforming to the design rules and estimated area for MDSA is 0.297 μm^2 whereas MDSANF occupies 0.5192 μm^2.

CCS Concepts: • **Hardware → Very large scale integration design**; **Process variations;**

Additional Key Words and Phrases: Low voltage SRAM, sensing delay, sense amplifier, SRAM sensing, transient analysis

ACM Reference format:
Ayush Dahiya, Poornima Mittal, and Rajesh Rohilla. 2023. Modified Decoupled Sense Amplifier with Improved Sensing Speed for Low-Voltage Differential SRAM. *ACM Trans. Des. Autom. Electron. Syst.* 28, 6, Article 94 (October 2023), 15 pages.
https://doi.org/10.1145/3611672

1 INTRODUCTION

The backbone of numerous system-on-chip (SoC) devices used in crucial industries like the **Internet of Things (IoT)**, biomedical implants, and wearable smart devices is a fast sensing, low-energy, and reliable **Static Random Access Memory (SRAM)**. The SRAM is typically the dominant contributor to power consumption as well area occupied of the whole chip [Zhu and Kursun 2014].

Authors' address: A. Dahiya, P. Mittal, and R. Rohilla, Department of Electronics and Communication Engineering, Delhi Technological University, Bawana Road, Shahbad Daulatpur Village, Rohini, New Delhi, Delhi, India, 110042; emails: ayushdahiya010@gmail.com, poornimamittal@dtu.ac.in, rajesh@dce.ac.in.
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
1084-4309/2023/10-ART94 $15.00
https://doi.org/10.1145/3611672

Realizing In-Memory Computing using Reliable Differential 8T SRAM for Improved Latency

AYUSH DAHIYA, Department of Electronics and Communication, Delhi Technological University, Delhi, India
POORNIMA MITTAL, Department of Electronics and Communication, Delhi Technological University, New Delhi, India
RAJESH ROHILLA, Department of Electronics and Communication, Delhi Technological University, Delhi, India

Traditional von Neumann computing architectures suffer from high energy and lower speed as compared to the requirements of modern applications like those required in neural network accelerators. A modified differential eight transistor (8^+T) static random access memory (SRAM)-based in-memory computing (IMC) structure was presented for realizing bit-wise Boolean logic operations. The 8^+T SRAM-IMC is designed at the 32 nm technology node with throughput of 2.1849, 2.4815, 2.5795, 2.6240, 2.6495, 2.6619, 2.6690, 2.6732, and 2.6749 giga outputs per second for 0.5 to 1.3 V supply voltage range, respectively. The differential 8T cell used to implement logic operations supports NAND and NOR operations with minimal overhead while also performing the standard storage operation with added stability over the conventional 6T and 8T SRAM cells. The SRAM-IMC offers reliable Boolean logic operations by using asymmetric sensing strategy for all process corners, TT, SS, SF, FS, and FF for an operating temperature range of 220 K to 400 K. Monte Carlo simulation considering global threshold voltage deviation of 50 mV was performed for various operating conditions to study the impact of variations on design parameters such as latency.

CCS Concepts: • **Hardware → Very large scale integration design**; **Process variations;**

Additional Key Words and Phrases: SRAM cell, in-memory computing, von Neumann bottleneck, latency, throughput

ACM Reference Format:
Ayush Dahiya, Poornima Mittal, and Rajesh Rohilla. 2024. Realizing In-Memory Computing using Reliable Differential 8T SRAM for Improved Latency. *ACM Trans. Des. Autom. Electron. Syst.* 29, 6, Article 106 (October 2024), 15 pages. https://doi.org/10.1145/3696666

1 Introduction

Technology scaling has paved the way for computing cost to be decreased to such an extent that we are surrounded by intelligent connected devices [Bai et al. 2022], but this constant struggle to make computing even faster, cost-efficient, and of lower power cannot continue perpetually.

Authors' Contact Information: Ayush Dahiya, Department of Electronics and Communication, Delhi Technological University, Delhi, Delhi, India; e-mail: ayushdahiya010@gmail.com; Poornima Mittal, Department of Electronics and Communication, Delhi Technological University, New Delhi, Delhi, India; e-mail: poornimamittal@dtu.ac.in; Rajesh Rohilla, Department of Electronics and Communication, Delhi Technological University, Delhi, Delhi, India; e-mail: rajesh@dce.ac.in.
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM 1084-4309/2024/10-ART106
https://doi.org/10.1145/3696666

INTEGRATION, the VLSI journal 106 (2026) 102586

Contents lists available at ScienceDirect

Integration, the VLSI Journal

journal homepage: www.elsevier.com/locate/vlsi

A low-power half-select free 8T SRAM cell with process-induced variation resistance for voltage scaling at 32 nm technology node

Ayush Dahiya, Poornima Mittal *, Rajesh Rohilla

Department of Electronics and Communication, Delhi Technological University, Delhi, India

ARTICLE INFO

Keywords:
Static random access memory (SRAM)
Threshold voltage mismatch
Variation tolerant circuits
Transient analysis
CMOS memory circuits
Multi-threshold CMOS

ABSTRACT

The SRAM cell structure in itself offers huge performance improvements in the memory array. An eight transistor SRAM cell (8TP) is presented with high-V_T PMOS transistors used together with nominal-V_T NMOS transistors. The 8TP cell offers a process variation resistant structure which offers a good balance between stability, delay, write margin, leakage and power consumption over other cell structures over a wide range of supply voltages for 0.6 V to 1.2 V. A number of 2000-point Monte Carlo simulations were performed for highlighting the impact of process, voltage and temperature variations on the cell architecture. The 8TP cell has a write-1/write-0 margin (WM) of 0.4091 V/ 0.4092 V at supply of 0.8 V. The stability of the cell is shown by its static noise margin (SNM) which shows a 6σ deviation for a temperature range of −25°C to 100°C. The cell offers excellent read and write performance compared to various other SRAM cells. The 8TP cell also demonstrates low leakage and power consumption compared among other 8T cells. The 8TP cell exhibits a good balance between speed and stability while having a competitive area footprint as well as half-select resilience which makes it a good choice for low-power applications.

1. Introduction

The relentless march of technology is driving a surge in smart devices and the widespread adoption of the Internet of Things (IoT) across diverse industries. This trend necessitates the development of increasingly energy-efficient architectures to power these advancements. One key application domain is wireless sensor networks (WSNs), which continuously monitor various environmental parameters using sensors and wireless transmission of data. However, this continuous operation comes at a significant energy cost, demanding extended battery life for these sensors [1–4]. The need for faster and less powerful devices has pushed digital designers into exploring deeper in the realm of SRAM [5,6]. Similarly, space applications like satellite communication, crucial for weather forecasting, military surveillance, and other critical functionalities, require low-cost, lightweight satellites with extended battery life and the ability to withstand harsh space radiation [7]. Some focus has also been put on the memory peripherals like the sense amplifier and their optimization in order to improve the memory array performance [8,9]. In these applications, high-density memories play a vital role, but they face the challenge of soft errors caused by high-energy particles like alpha (α) and beta (β) particles that can get trapped within semiconductor devices, corrupting stored data and jeopardizing circuit performance. Addressing these challenges, recent research in SRAM bitcell design has explored the potential of bit-interleaving architectures as a promising solution to mitigate soft errors. This technique offers a compelling approach to enhance the reliability of memory devices in these demanding applications.

Over the years, the literature has introduced various configurations of SRAM cells with the aim of achieving superior performance compared to conventional designs [10,11]. Standard 8T SRAM cell improves Read Stability Noise Margin by employing isolated read path through the inclusion of two additional transistors. However, this benefit comes with a trade-off of increased read path leakage, which worsens with technology miniaturization. To address this, [12] proposes a modified isolated read path design within their SRAM cell. This modification specifically tackles data-independent leakage, thereby mitigating overall leakage and consequently improving RSNM. [13] introduced a novel SRAM cell design that integrates a Schmitt-trigger inverter to enhance both read and write stability margins. This innovative approach capitalizes on the robust noise margin properties inherent in the voltage transfer curve of the Schmitt-trigger, thereby improving noise immunity. However, it is worth noting that the inclusion of the Schmitt-trigger INV leads to a larger cell footprint. Multiple studies [14–17] have explored the adoption of single bitline architectures in SRAM cells as a strategy to lower power consumption. This approach centers around reducing the bitline switching activity factor (α) below 50%

* Corresponding author.
E-mail address: poornimamittal@dtu.ac.in (P. Mittal).

https://doi.org/10.1016/j.vlsi.2025.102586
Received 14 March 2025; Received in revised form 29 September 2025; Accepted 17 October 2025
Available online 1 November 2025
0167-9260/© 2025 Published by Elsevier B.V.

Comparative Analysis of Current Sense Amplifier Architectures for SRAM at 45 nm Technology Node



Ayush, Poornima Mittal, and Rajesh Rohilla

Abstract Static random access memory is an essential part of digital system-on-chip-based designs. The advanced scaling of devices has put forth an increased focus on the power requirements of memories and its peripheral devices with sense amplifier (SA) being a key component of their read operation. This results in the selection of a SA architecture becoming vital for improved speed defined in terms of delay and reduced power consumption. A comparative analysis of four current sense amplifiers, namely clamped bit line sense amplifier (CBLSA), hybrid current sense amplifier (HCSA), modified current sense amplifier (MCSA) and modified hybrid sense amplifier (MHSA) for 45 nm technology, has been presented. The delay and power consumption for the four architectures have been depicted at varied supply voltages, and their consequent behavior is analyzed. This paper aims to provide an analysis of the current sense amplifiers architectures applicable as a selection criterion for SA in SRAM design.

Keywords SRAM · Current sense amplifier · Current conveyer

1 Introduction

The modern computational devices are evolving in terms of their sizes, operations as well as consumer expectations. Most of these devices require some form of embedded memory. The advances in CMOS scaling have lent themselves to being an impacting factor in the design of highly dense architectures of memory that provide better performance while also maintaining system reliability. Thus, the memory is a primary component of the computation chain both in terms of delay and power consumption.

SRAM stands for static random access memory, and it is one of the key memory types for modern low-power applications. SRAM memories are designed in the form

Ayush (✉) · P. Mittal · R. Rohilla
Department of Electronics and Communications Engineering, Delhi Technological University, Delhi 110042, India
e-mail: ayushdahiya010@gmail.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

S. Das et al. (eds.), *Advances in Data-Driven Computing and Intelligent Systems*,
Lecture Notes in Networks and Systems 698,
https://doi.org/10.1007/978-981-99-3250-4_48



Design Trade-Offs in 8T SRAM Architectures: Differential and Single-Ended Approaches

Ayush Dahiya(✉), Poornima Mittal, and Rajesh Rohilla

Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India
ayushdahiya010@gmail.com, poornimamittal@dtu.ac.in, rajesh@dce.ac.in

Abstract. This work presents a comparative evaluation of differential (8TD) and single-ended (8TS) 8T SRAM cell architectures designed using 32 nm CMOS technology, across a supply voltage range of 0.6 V to 1.2 V. At near-threshold operation ($V_{DD} = 0.6$ V), the 8TD cell demonstrates a static noise margin (SNM) of 0.1932 V—representing a 26.7% improvement over the 8TS cell, which achieves 0.1525 V. Monte Carlo simulations with 2000 samples further confirm the 8TD cell's robustness, exhibiting a standard deviation of only 5.6 mV in SNM, compared to 15.2 mV for the 8TS design—a reduction of over 60%. In terms of speed, the 8TD achieves a read delay of 340 ps versus 400 ps for the 8TS at 0.6 V, and maintains consistent write delays between 110–125 ps across voltage scaling. The 8TS cell, on the other hand, exhibits greater asymmetry and degradation in write performance at low voltages, particularly under write margin 0 conditions, where the access transistor connected to a low internal node must flip the cell. While the 8TD architecture involves additional control complexity and higher dynamic power due to differential sensing, it offers significantly better stability, variation tolerance, and access speed under low-voltage operation—making it a strong candidate for energy-efficient, variation-resilient memory systems.

Keywords: Low Voltage SRAM · Latency · Voltage scaling · Variation tolerant circuits · Static Noise Margin

1 Introduction

The continuous scaling of CMOS technologies for reducing area footprint and enhancing performance in advanced computing applications, particularly artificial intelligence (AI) accelerators, has heightened the vulnerability of integrated circuits to variability-induced reliability issues. In deeply scaled technology nodes, SRAM cells–which typically dominate transistor count and die area in contemporary systems-on-chips (SoCs)–experience intensified reliability challenges due to front-end wearout mechanisms, such as bias temperature instability

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
S. J. Nanda et al. (Eds.): ICDSA 2025, LNNS 1795, pp. 123–133, 2026.
https://doi.org/10.1007/978-3-032-15404-0_11

International Conference on Advances in Data-driven Computing and Intelligent Systems (ADCIS 2022)

BITS Pilani
K K Birla Goa Campus
Center for Technical Education

Organized by
BITS Pilani, K K Birla Goa Campus, India
Technically Sponsored by
Soft Computing Research Society
September 23-25, 2022

SOFT COMPUTING RESEARCH SOCIETY
S C R S
सा विद्या या विमुक्तये

Certificate Of Participation

This is to certify that **Ayush** has presented the paper titled "**Comparative Analysis of Current Sense Amplifiers Architectures for SRAM at 45 nm Technology Node**" authored by **Ayush, Poornima Mittal and Rajesh Rohilla** in the International Conference on Advances in Data-driven Computing and Intelligent Systems (ADCIS 2022) held during September 23-25, 2022.

SCRS\ADCIS2022\399

Dr. Swagatam Das
General Chair

Dr. Snehanshu Saha
General Chair

Dr. Jagdish Chand Bansal
General Chair

6th International Conference on Data Science and Applications (ICDSA 2025)

Organised By

Malaviya National Institute of Technology Jaipur, India

Technically Sponsored by

Soft Computing Research Society

CERTIFICATE

OF PRESENTATION

This certificate is awarded to

Ayush Dahiya

who has presented the paper titled **"Design Trade-Offs in 8T SRAM Architectures: Differential and Single-Ended Approaches"** authored by **Ayush Dahiya, Poornima Mittal, Rajesh Rohilla** in the **6th International Conference on Data Science and Applications (ICDSA 2025)** held at **Malaviya National Institute of Technology Jaipur, India** during **July 16-18, 2025.**

SCRS\ICDSA2025\PC\1835

SJNanda

Dr. Satyasai Jagannath Nanda
Malaviya National Institute of Technology Jaipur
(General Chair)

Prof. Rajendra Prasad Yadav
Malaviya National Institute of Technology Jaipur
(General Chair)

Springer

https://scrs.in/conference/icdsa2025

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis: **SRAM Design for Throughput Enhancement**

Total Pages: **175** Name of the Scholar: **Mr. Ayush**

Supervisor (s):

(1) **Prof. Poornima Mittal,** Professor, Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India

(2) **Prof. Rajesh Rohilla,** Professor, Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India

Department: **Electronics and Communication Engineering**, Delhi Technological University

This is to report that the above thesis was scanned for similarity detection. Process and outcome are given below:

Software used: **Turnitin** Similarity Index: **8%** Total Word Count: **62,282**



Candidate's Signature



Signature of Joint Supervisor



Signature of Supervisor

Poornima Mittal

PhD Thesis_2K21PHDEC19_Ayush

btech project

Document Details

Submission ID
trn:oid:::27535:139751820

Submission Date
May 20, 2026, 8:13 PM GMT+5:30

Download Date
May 20, 2026, 8:21 PM GMT+5:30

File Name
PhD Thesis_2K21PHDEC19_Ayush.pdf

File Size
43.5 MB

175 Pages

62,282 Words

344,685 Characters



8% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 8 words)

Exclusions

- 6 Excluded Sources

Match Groups

399 Not Cited or Quoted 8%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

3% Internet sources
6% Publications
3% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **399** Not Cited or Quoted 8%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

- 3% Internet sources
- 6% Publications
- 3% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Publication	Arunima Tripathi, Ayush Dahiya, Poornima Mittal. "A Low-Power Single Ended Hal... 1%
2	Publication	Ayush Dahiya, Vansh Singhal, Poornima Mittal. "A Variation Tolerant Write Assist ... <1%
3	Student papers	Delhi Technological University on 2026-04-30 <1%
4	Internet	ebin.pub <1%
5	Internet	iris.unipv.it <1%
6	Student papers	Anna University on 2025-12-26 <1%
7	Publication	Ceriz, Ernesto Jorge Ferreira. "O rasto, o indício, o vestígio, como fontes de criaçã... <1%
8	Internet	arxiv.org <1%
9	Internet	dspace.iiti.ac.in <1%
10	Publication	Arunima Tripathi, Ayush Dahiya, Poornima Mittal. "A low-power single ended half... <1%

11	Publication	Zhu, Hong. "Power Efficiency and Reliability Enhancement Techniques in SRAM Ci... <1%
12	Internet	s3-eu-west-1.amazonaws.com <1%
13	Publication	Mahammad Aadil S, Bhargava Rama M, Yashwant K S S, Latha P. "Design of stable... <1%
14	Internet	mospace.umsystem.edu <1%
15	Student papers	Uttar Pradesh Technical University on 2025-07-01 <1%
16	Internet	tudr.thapar.edu <1%
17	Internet	publikationen.bibliothek.kit.edu <1%
18	Publication	Ming-Bo Lin. "Introduction to VLSI Systems - A Logic, Circuit, and System Perspect... <1%
19	Student papers	University of Missouri, Kansas City on 2020-06-15 <1%
20	Publication	Bhawna Rawat, Poornima Mittal. "A latch-based sense amplifier with improved p... <1%
21	Publication	Yin, Zihan. "Analog In-Pixel Computing for Energy-Efficient Visual Intelligence at t... <1%
22	Publication	Chandramauleshwar Roy, Aminul Islam. "Design of low power, variation tolerant ... <1%
23	Student papers	ABV-Indian Institute of Information Technology and Management Gwalior on 201... <1%
24	Student papers	Universiteit Hasselt on 2025-06-09 <1%

25 Publication
Jahinuzzaman, Shah M. "Modeling and mitigation of soft errors in nanoscale SRA... <1%

26 Student papers
Rochester Institute of Technology on 2014-09-18 <1%

27 Student papers
M S Ramaiah University of Applied Sciences on 2025-01-06 <1%

28 Internet
impa.usc.edu <1%

29 Student papers
Delhi Technological University on 2026-04-12 <1%

30 Internet
d197for5662m48.cloudfront.net <1%

31 Publication
Salahuddin, Shairfe Muhammad. "Low-Power Embedded Memory Circuits in Nan... <1%

32 Publication
"Sub-threshold Design for Ultra Low-Power Systems", Springer Science and Busin... <1%

33 Student papers
Anna University on 2024-12-12 <1%

34 Publication
Bhawna Rawat, Poornima Mittal. "A 32 nm single ended single port 7T SRAM for l... <1%

35 Publication
Elakkumanan, Praveen. "Overcoming the circuit design challenges in nanoscale S... <1%

36 Publication
Ishfag Ahmad, Sanjay Ranka. "Handbook of Energy-Aware and Green Computing,... <1%

37 Publication
Jamalapurapu, Sri Vaishnavi. "Characterization of Sneak Path Current Effect in a P... <1%

38 Publication
James E. Morris, Krzysztof Iniewski. "Nanoelectronic Device Applications Handbo... <1%

39 Internet
faculty.coe.drexel.edu <1%

40 Internet
www.biologicalengineering.in <1%

41 Publication
Anil Kumar Rajput, Manisha Pattanaik, Gaurav Kaushal. "An energy-efficient 10T ... <1%

42 Publication
Erfan Abbasian, Shilpi Birla, Morteza Gholipour. "Ultra-low-power and stable 10-n... <1%

43 Publication
Refaei, Akram Mohamed Ahmed Araby. "Implementing a Reliable and Ultra-low P... <1%

44 Internet
infoscience.epfl.ch <1%

45 Publication
Amir Dolatshah, Erfan Abbasian, Maryam Nayeri, Sobhan Sofimowloodi. "A sub-th... <1%

46 Student papers
International Institute of Information Technology, Hyderabad on 2026-05-02 <1%

47 Internet
hdl.handle.net <1%

48 Publication
Zhang, Qirui. "Domain-Specific Acceleration: From Efficient Vision Processing Har... <1%

49 Publication
Carles Hernández Luz. "Addressing Manufacturing Challenges in NoC-based ULSI ... <1%

50 Student papers
Indian Institute of Technology Indore on 2026-03-20 <1%

51 Student papers
JSS Science & Technology University, Mysore on 2025-06-05 <1%

52 Student papers
Anna University on 2024-10-05 <1%

53 Student papers
Delhi Technological University on 2026-04-11 <1%

54 Publication
Shams ul Haq, Erfan Abbasian, Tabassum Khurshid. "A Power/Energy-Efficient 8T ... <1%

55 Publication
Xin Li, Mengya Gao, Zihua Ren, Kefeng Yu, Wenjuan Lu, Chenghu Dai, Wei Hu, Ch... <1%

56 Internet
os.unil.cloud.switch.ch <1%

57 Internet
vtechworks.lib.vt.edu <1%

58 Publication
Ali, Mustafa. "Energy-Efficient Deep Learning Hardware Design Through Coupling... <1%

59 Publication
Anil Kumar Rajput, Manisha Pattanaik, Gaurav Kaushal. "Local bit-line shared pas... <1%

60 Publication
Anil Rajput, Manisha Pattanaik, Gaurav Kaushal, Santosh Kumar Vishvakarma. "A ... <1%

61 Student papers
Anna University on 2025-07-30 <1%

62 Publication
Tavva, Sreeharsha. "A novel variation-tolerant 9T SRAM design for nanoscale CM... <1%

63 Student papers
Universiti Malaysia Perlis on 2017-05-30 <1%

64 Internet
gyan.iitg.ac.in <1%

65 Student papers
Anna University on 2025-12-13 <1%

66 Publication
Elangovan Mani, Padmaja Nimmagadda, Shaik Javid Basha, Mohammed A. El-Meli... <1%

67 Student papers
Glyndwr University on 2014-04-18 <1%

68 Publication
Thirumala, Sandeep Krishna. "Enabling Logic-Memory Synergy Using Integrated ... <1%

69 Publication
Vansh Singhal, Bhawna Rawat, Poornima Mittal, Brijesh Kumar. "A single ended, ... <1%

70 Publication
Venu Birudu, Tirumala Rao Kadiyam, Koteswara Rao Penumalli, Siva Sankar Yella... <1%

71 Internet
dspace.thapar.edu:8080 <1%

72 Internet
www.ideals.illinois.edu <1%

73 Internet
www.researchgate.net <1%

74 Student papers
Anna University on 2024-09-12 <1%

75 Student papers
University of Edinburgh on 2023-08-21 <1%

76 Publication
Zhiting Lin, Runru Yu, Da Huo, Qingchuan Zhu et al. "A 28-nm 9T SRAM-based CIM... <1%

77 Publication
Kulkarni, Jaydeep Prakash. "Low voltage robust memory circuit design", Proquest... <1%

78 Publication
"IEEE-ICDCS conference proceeding", 2012 International Conference on Devices Ci... <1%

79 Publication
Charles, Subodha. "Design of Secure and Trustworthy Network-on-Chip Architect... <1%

80 Publication
John D. Cressler, H. Alan Mantooth. "Extreme Environment Electronics", CRC Pres... <1%

81 Publication
Pan Yang, Xiaocan Ye, Yongxin Zhao, Wei Zhang, Shoumou Huang, Yang Huang, Y... <1%

82 Student papers
RMIT University on 2022-09-19 <1%

83 Internet
theses.lib.polyu.edu.hk <1%

84 Publication
Aayush Ankit, Indranil Chakraborty, Amogh Agrawal, Mustafa Ali, Kaushik Roy. "C... <1%

85 Publication
Engineering Computations, Volume 31, Issue 2 (2014-03-28) <1%

86 Student papers
Indian Institute of Technology Roorkee on 2025-08-18 <1%

87 Publication
Ogbogu, Chukwufumnanya Osazee. "High Performance and Reliable Processing-I... <1%

88 Student papers
Rijksuniversiteit Groningen - Tii on 2026-04-15 <1%

89 Publication
Seyed Hassan Hadi Nemati, Nima Eslami, Mohammad Hossein Moaiyeri. "A Hybri... <1%

90 Internet
theses.hal.science <1%

91 Publication
"Proceeding of the Second International Conference on Microelectronics, Comput... <1%

92 Student papers
AUT University on 2024-10-21 <1%

93 Publication
Antoni Roca Pérez. "Floorplan-Aware High Performance NoC Design", Universitat ... <1%

94 Publication
Bhawna Rawat, Poornima Mittal. "A single–Port to Dual-Port Reconfigurable 7T S... <1%

95 Publication
Biby Joseph, R.K. Kavitha. "A radiation hardened In-memory computing SRAM for ... <1%

96 Publication
Cho, Minchang. "Distributed, Intelligent Audio Sensing Enabled by Low-Power Int... <1%

97 Publication
Eckert, Charles S.. "In-SRAM Computing for Neural Network Acceleration", Univer... <1%

98 Publication
Farzaneh Izadinasab, Morteza Gholipour. "Half-select disturb-free single-ended 9-... <1%

99 Publication
Luo, Yandong. "Energy-Efficient On-Chip Deep Neural Network (DNN) Inference a... <1%

100 Publication
Shrimali, Kaustubh. "Surface Chemistry Considerations in the Reverse Flotation o... <1%

101 Student papers
University of Glasgow on 2023-04-27 <1%

102 Student papers
University of Glasgow on 2024-04-24 <1%

103 Student papers
University of Macau on 2025-11-04 <1%

104 Student papers
Uttar Pradesh Technical University on 2020-11-04 <1%

105 Publication
Vishal Sharma, Hyunjoon Kim, Tony Tae-Hyoung Kim. "A 64 Kb Reconfigurable Ful... <1%

106 Publication
Wai-Kai Chen. "Memory, Microprocessor, and ASIC", CRC Press, 2019 <1%

107 Internet
journals.uob.edu.bh <1%

108 Publication
Alireza Abbasi, Farbod Setoudeh, Mohammad Bagher Tavakoli, Ashkan Horri. "A ... <1%

109 Publication
Bhawna Rawat, Poornima Mittal. "A Switching NMOS Based Single Ended Sense A... <1%

110 Publication
Bianchi, Carlo. "Dependence of Building Energy Consumption and Solar PV Perfor... <1%

111 Student papers
Birla Institute of Technology on 2018-07-07 <1%

112 Student papers
Birla Institute of Technology on 2018-07-22 <1%

113 Publication
Chuan-Jia Jhang, Cheng-Xin Xue, Je-Min Hung, Fu-Chun Chang, Meng-Fan Chang. "... <1%

114 Publication
Erfan Abbasian, Shams ul Haq, Pavan Kumar Mukku, shilpi birla. "Comparative Ev... <1%

115 Publication
Erfan Abbasian, Tahere Mirzaei, Sobhan sofimowloodi. "A Stable Low Leakage Po... <1%

116 Publication
G. Ramesh, T. Venkatakrishnamoorthy, Madhusudhana Reddy Barusu, M.K. Linga... <1%

117 Publication
Hanoi Pedagogical University 2 <1%

118 Publication
Hossein Khosravi, Ricardo Carmona Galán, Jorge Fernández-Berni, Nabeeh Kanda... <1%

119 Student papers
Indian Institute of Technology on 2020-01-09 <1%

120 Student papers
Indian Institute of Technology, Kanpur on 2024-05-15 <1%

121 Student papers
Indian Institute of Technology, Kanpur on 2024-06-21 <1%

122 Student papers
Indian School of Mines on 2019-04-29 <1%

123 Publication
Kodge, Sangamesh. "Towards Efficient and Robust Deep Learning: Handling Data ... <1%

124 Student papers
M S Ramaiah University of Applied Sciences on 2025-02-12 <1%

125 Publication
M. Elangovan, Kulbhushan Sharma, Haitham A. Mahmoud, Ashish Sachdeva, S. Je... <1%

126 Publication
Madan Mohan Sharma, Ananya Kabba, Kulbhushan Sharma, Pankaj Kumar. "High... <1%

127 Student papers
Malaviya National Institute of Technology on 2020-07-01 <1%

128 Publication
Meng, Jian. "Efficient, Robust, and Self-Supervised Algorithm-Hardware Co-Design... <1%

129 Publication
N. Lepri, A. Glukhov, L. Cattaneo, M. Farronato, P. Mannocci, D. Ielmini. "In-memo... <1%

130 Student papers
National Sun Yat-sen University on 2023-11-21 <1%

131 Publication
Odom, Gaines. "Leveraging SRAM For Counterfeit Detection and Secure 3DIC Inte... <1%

132 Publication
Paul, Partha Sarathi. "Enhanced-Entropy Chaotic Circuit Design for Overhead-Con... <1%

133 Publication
Rupesh D. Kadhao, Siddharth R. K., Nithin Kumar Y. B., Vasantha M. H., Devesh D... <1%

134 Publication
Sharma, Harsh. "Ai-Driven Design and Optimization of Heterogeneous Chiplet Sys... <1%

135 Student papers
V.B.S Purvanchal University, Jaunpur on 2024-04-01 <1%

136 Publication
Venu Birudu, Siva Sankar Yellampalli, Ramesh Vaddi. "A negative capacitance FET ... <1%

137 Publication

Vishal Sharma, Ju-Eon Kim, Hyunjoon Kim, Lu Lu, Tony Tae-Hyoung Kim. "A Recon... <1%

138 Publication

Yue Zhao, Jinkai Wang, Zhongzhen Tong, Xiulong Wu, Chunyu Peng, Wenjuan Lu, ... <1%

139 Publication

Yufei Ma, Yuan Du, Li Du, Jun Lin, Zhongfeng Wang. "In-Memory Computing: The ... <1%

140 Internet

ndl.ethernet.edu.et <1%

141 Internet

websrv.cecs.uci.edu <1%



AYUSH

+91-9466040555 | ayushdahiya010@gmail.com
ORCID 0000-0002-5196-231X

Employment

Assistant Professor
Galgotias College of Engineering and Technology
Department of Electronics and Communication Engineering

Education

Ph.D. in Electronics and Communication Engineering | 2026
Delhi Technological University, Delhi

M.Tech. in Electronics and Communication Engineering | 2021
National Institute of Technology, Delhi

B.Tech. in Electronics and Communication Engineering | 2019
University Institute of Engineering & Technology, MDU Rohtak

Journal Publications

1. **A. Dahiya**, P. Mittal and R. Rohilla (2025). A Low-Power Half-Select Free 8T SRAM Cell with Process-Induced Variation Resistance for Voltage Scaling at 32 nm Technology Node. *Integration, 106, 102586.*
2. **A. Dahiya**, V. Singhal and P. Mittal (2025). A Variation Tolerant Write Assist Read Decoupled 9T SRAM Cell for Low Voltage Application. *ACM TODAES, 30(5), Article 82.*
3. A. Tripathi, **A. Dahiya** and P. Mittal (2024). A Low-Power Single Ended Half-Select Free 7T SRAM Cell with Improved Write Margin at 32 nm Technology Node. *Physica Scripta, 100(1), 015015.*
4. **A. Dahiya**, P. Mittal and R. Rohilla (2024). Realizing In-Memory Computing Using Reliable Differential 8T SRAM for Improved Latency. *ACM TODAES, 29(6), 1-15.*
5. **A. Dahiya**, P. Mittal and R. Rohilla (2023). Modified Decoupled Sense Amplifier with Improved Sensing Speed for Low-Voltage Differential SRAM. *ACM TODAES, 28(6), Article 94.*
6. **A. Dahiya**, P. Mittal and R. Rohilla (2023). Variation-Tolerant Sense Amplifier Using Decoupling Transistors for Enhanced SRAM Read Performance. *Circuits, Systems, and Signal Processing, 42(10), 5799-5810.*

Conference Publications

1. **A. Dahiya**, P. Mittal and R. Rohilla (2026). Design Trade-Offs in 8T SRAM Architectures: Differential and Single-Ended Approaches. *ICDSA 2025, Springer LNNS 1795, 123-133.*
2. **A. Dahiya**, A. Tripathi, D. Usham and P. Mittal (2024). A Low Power Single Ended and Dual Port 8T SRAM with Robust Features. *CISCON 2023, Springer LNEE 1236, 301-316.*
3. **A. Dahiya**, P. Mittal and R. Rohilla (2023). Comparative Analysis of Current Sense Amplifier Architectures for SRAM at 45 nm Technology Node. *ADCIS 2022, Springer LNNS 698, 633-640.*

Technical Skills

Research SRAM and memory design, sense amplifiers, SRAM in-memory computing, CMOS circuits.
Programming C, MATLAB, Verilog, Perl.
EDA tools Cadence Design Suite, Xilinx Vivado, LTspice; open-source RTL-to-GDSII workflows.
Systems Linux, Red Hat Enterprise Linux, Ubuntu; scripting and tool integration.