Design of minimum energy driven ultra-low voltage SRAMs and D flip-flop

Wang, Bo

2015

Wang, B. (2015). Design of minimum energy driven ultra-low voltage SRAMs and D flip-flop.
Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/65355>

<https://doi.org/10.32657/10356/65355>

**Design of Minimum Energy Driven
Ultra-low Voltage
SRAMs and D Flip-Flop**

WANG BO

**SCHOOL OF ELECTRICAL AND ELECTRONIC
ENGINEERING**

2015

**Design of Minimum Energy Driven
Ultra-Low Voltage
SRAMs and D Flip-Flop**

WANG BO

School of Electrical and Electronic Engineering

A thesis submitted to the Nanyang Technological University
in partially fulfillment of the requirement for the degree of
Doctor of Philosophy
2015

Acknowledgement

My first and sincere gratitude goes to my supervisor, Prof. Tony Tae-Hyoung Kim, for his continuous guidance and significant support in my PhD life. His well-planned research perspective extremely helps me to establish the research goals and develop the research path. His enthusiasm in training students provides me opportunities to discuss with him face to face every time I need and enormously inspires me. His generosity in financial support endows me many chances to attend international conferences and broadens my views. Without his enlightenment and encouragement, I would never have been able to achieve any tiny progress in research.

I would like to thank my co-supervisor, Dr Zhou Jun, for his devotion in my PhD research. He offers me a chance to work with him and learn from him. He is very open and ready to help either in idea sharing or career planning, which benefits me immensely. I would also give my appreciation to Prof. Yeo Kiat Seng, Prof. See Kye Yak, Prof. Siek Liter, Prof. Goh Wang Ling, Prof. Zhang Yue Ping, Prof. Boon Chirn Chye, Prof. Chan Pak Kwong, Prof. Lam Ying Hung, Prof. Kong Zhi Hui, Prof. Zheng Yuanjing, Dr Tan Khen Sang and all the technical staffs in the VIRTUS lab, the VLSI lab and the IC Design lab.

I want to give my thanks to my colleagues both in NTU and in the Institute of Microelectronics, A*STAR for their enormous help. They are Prof. Je Minkyu, Dr Liu Xin, Dr Wang Chao, Dr Hylas Lam, Dr Do Ahn Tuan, Dr Mohammed Sultan Mohiuddin Siddiqui, Liu Lizhuang, Chang Kah Hyong, Lan Jingjing, Lim Ching Yun, Lee Zhao Chuan, Seyed Mohammad Ali Zeinolabedin, Aung Myat Thu Linn, Le Ba Ngoc, Karim Hany Mohamed Rawy, Abhik Das, Neelakantan Narasimman, J. Karthik Gopal, Kavitha Velayudhan, Achiranshu Garg, Qi Li, Truc Quynh Nguyen and Yeo Yuan Lin. I am very pleased to work with them and be friends with them.

My thanks also go to Dr Chong Sau Siong, Dr He Xiaofeng, Dr Li Sizhen, Dr Lu Zhenghao, Dr Chris Yeung, Dr Huang Xiwei, Dr Fei Wei, Dr Jeremy Low Yung Shern, Dr Joshua Low Yung Lih, Dr Yu Jun, Dr Liu Chang, Dr Li Yan, Ms Yang

Wanlan, Howard Tang, Zou Qiong, Ye Wanxin, Zhu Yao, Yang Yan, Chen Yi, Cai Deyun, Chen Zihao, Wang Yanmei, Xu Shanshan, Meng Fanyi, Lu Lu, Zhang Ying, Sun Junyi, Feng Guangyin, Huang Nan, Feng Xiaohua, Yao Enyi, Qiu Lei, Tay Thian Fatt, Zhang Le, Wang Yong, Deng Tianwei, Wu Chundong, Yang Yongkui, Han Beibei, Yi Xiang, Zhang Xiangyu, Qian Xinyuan, Yu Hang, Chen Yi, Tan Xiaoliang, and Zhao Jianming for their sharing and helping.

Last but not least, I give my heartiest thanks to my parents who always support me with love. They create a world for me and guide me to explore a bigger one. They are the greatest mom and dad.

Table of Contents

Acknowledgement.....	I
List of Figures	VII
List of Tables	XIII
Chapter 1 Introduction	1
1.1 Motivation	1
1.2 Research Objectives and Contributions.....	4
1.3 Organizations.....	7
Chapter 2 Background and Literature Review	8
2.1 Conventional 6T Single-port SRAMs	8
2.1.1 6T Single-port SRAM Operation	8
2.1.2 Challenges of 6T SRAMs for Ultra-low Voltage Operation	11
2.1.3 Design Techniques of 6T SRAMs to Improve Minimum Voltage	15
2.2 Conventional 8T Dual-port SRAMs.....	18
2.2.1 8T Dual-port SRAM Operation.....	19
2.2.2 Challenges of 8T SRAMs for Ultra-low Voltage Operation	21
2.2.3 Design Techniques of 8T SRAMs with Ultra-low Supply Voltage....	23
2.3 Conventional D Flip-Flop Circuits	27
2.3.1 Mainstream DFF Circuits and Timing Properties	27
2.3.2 Design Challenges of DFFs for Energy Efficient Applications	30

2.3.3	Design Techniques for Energy Efficient DFFs.....	31
Chapter 3	SRAM Device and Circuits Optimization toward Energy Efficiency in Multi- V_{th} CMOS	34
3.1	Background.....	34
3.2	Analysis of SRAM Energy	36
3.2.1	SRAM Energy Modeling.....	36
3.2.2	Effects of Supply Voltage Scaling and Threshold Voltage on Energy Efficiency	39
3.2.3	Effects of Multi- V_{th} Devices on SRAM Energy	40
3.3	Minimum Energy-Driven SRAM Design Utilizing Multi- V_{th} Devices	42
3.3.1	Analysis of SRAM Energy without Multi- V_{th} Devices	43
3.3.2	Analysis of SRAM Energy with Multi- V_{th} Devices.....	44
3.4	Design Techniques for SRAM Energy Efficiency Improvement Utilizing Multi- V_{th} Devices.....	49
3.4.1	Effect of Power Reduction Techniques on SRAM Energy	50
3.4.2	Effect of Performance Boosting Techniques on SRAM Energy	52
3.4.3	Combination Effect of Power Reduction and Performance Boosting Techniques.....	55
3.5	Summary.....	57
Chapter 4	Design of an Ultra-low Voltage 9T SRAM with Equalized Bitline Leakage and CAM-assisted Energy Efficiency.....	58
4.1	Background.....	58
4.2	Proposed SRAM Design Techniques for Ultra-low Voltage Operation	60

4.2.1	A Novel 9T SRAM Cell	60
4.2.2	Analysis of Static Noise Margin and Write Margin	62
4.2.3	Bitline Leakage Equalization with the Worst Case of Leakage	64
4.3	Proposed Energy Efficient Improvement Technique	69
4.3.1	Limitation of MTCMOS on SRAM Energy Efficiency	69
4.3.2	Proposed CAM-assisted Write Performance Boosting Technique	72
4.4	Test Chip Implementation and Measurement	78
4.5	Summary	82
Chapter 5	Design of an Ultra-low Voltage Disturb-suppressed Dual-port SRAM	83
5.1	Background	83
5.2	Proposed 12T DP SRAM Cell	85
5.2.1	12T SRAM Cell Design	86
5.2.2	Implementation of Virtual Ground for Bitline Leakage Reduction ...	87
5.3	Disturb Suppression of 12T DP SRAM in Common-Row-Access Mode	89
5.3.1	Analysis of Disturb Occurrence Probability	89
5.3.2	Analysis of Read Disturb	91
5.3.3	Analysis of Write Disturb	94
5.4	Measurement Results	96
5.5	Summary	99
Chapter 6	Design of an Ultra-low Voltage, Energy-Delay Efficient Charge-Pumped DFF	100

6.1	Background.....	100
6.2	A Novel Sub-threshold DFF	103
6.2.1	DFF Circuit Design and Near-/Sub-threshold Operation.....	103
6.2.2	Inverse-Narrow-Width-Effect-Aware Sizing Strategy	106
6.3	Analysis of CPDFF with TGFF and ACFF	108
6.3.1	C-Q Delay Investigation.....	108
6.3.2	Comparison of Setup Time and Hold Time.....	111
6.3.3	Analysis of Energy-Delay Product.....	113
6.4	Test Chip Implementation and Measurement	114
6.5	Summary.....	118
Chapter 7	Conclusions and Future Works.....	119
7.1	Conclusions	119
7.2	Future Works	120
Publications		121
Bibliography		123

List of Figures

Fig. 1.1 Process feature size trend [1].	1
Fig. 1.2 Power dissipation as a function of VDD for the 16 kb 9T SRAM [6].....	2
Fig. 1.3 Energy dissipation as a function of VDD for the 16b 1024-point FFT [7]...	3
Fig. 2.1 Schematic of conventional 6T SRAM cell.	8
Fig. 2.2(a) Concept of read operation in a 6T SRAM bitline. (b) Concept of write operation in a 6T SRAM bitline.	9
Fig. 2.3 Cell stability degradation of 6T SRAM cell due to read disturb.	11
Fig. 2.4 Illustration of SNM of 6T SRAM cell.	12
Fig. 2.5(a) 6T SRAM read SNM by VDD sweeping. (b) 6T SRAM write margin by VDD sweeping.	14
Fig. 2.6(a) Schematic of 8T SRAM cell [13]. (b) Schematic of 10T SRAM cell [17].	16
Fig. 2.7(a) Schematic of conventional 8T dual-port SRAM. (b) Parallel memory access of 8T dual-port SRAM [30].	18
Fig. 2.8(a) Illustration of different-row-different-column access. (b) Illustration of different-row-same-column access. (c) Illustration of common-row-different-column access. (d) Illustration of common-row-common-column access [30].....	19
Fig. 2.9 Comparison of read SNMs in different-row access and common-row access situations [30].....	22
Fig. 2.10 Concept of access circumvention scheme for dual-port 8T SRAM [30]..	23
Fig. 2.11 Concept of active bitline equalizing technique for dual-port 8T SRAM [32].	24
Fig. 2.12(a) Write-disturb detector for 8T dual-port SRAM. (b) Coordinately-	

activated write drivers for dual-port SRAM [32].	25
Fig. 2.13 Schematic of transmission-gate FF (TGFF).	27
Fig. 2.14 Schematic of True-Single-Phase-Clocked (TSPC) FF.	28
Fig. 2.15 Illustrating DFF setup time and hold time [34].	29
Fig. 2.16 Schematic of adaptive-coupling FF (ACFF) [37].	31
Fig. 2.17 Schematic of Static-Single-Phase Contention-Free FF (S^2CFF) and its operation [38].	33
Fig. 3.1 Simplified SRAM array diagram for energy analysis.	36
Fig. 3.2 Schematic of an 8T decoupled SRAM cell with multi- V_{th} devices.	41
Fig. 3.3 Normalized energy of three SRAMs designed by three different device types (i.e. HVT, SVT and LVT). All transistors in one SRAM have the same V_{th} .	43
Fig. 3.4 Impact of device selection on normalized energy of three SRAMs. Note that HVT devices are employed for read port in all SRAMs. Rest transistors in each SRAM cell adopt one device type.	44
Fig. 3.5 Normalized delay values of SRAM read and write operations designed with HVT devices.	44
Fig. 3.6 Comparison of read delay (LVT) with write delay implemented with multi- V_{th} devices (SVT and HVT).	45
Fig. 3.7 Normalized energy of SRAMs utilizing three different device types (i.e. HVT, SVT and LVT) for data storage and write paths. Note that LVT devices are used in read port.	46
Fig. 3.8 Comparison of leakage current over various device combinations.	46
Fig. 3.9 Summary of normalized minimum energy consumption over various device combinations.	47
Fig. 3.10 Summary of normalized leakage current over various device combinations.	

.....	47
Fig. 3.11 8T decoupled SRAM cells with leakage reduction techniques: (a) column-interleaved and (b) read buffer foot control	49
Fig. 3.12 Effect of column-interleaved scheme on SRAM energy. The reference design is using SVT devices in the write paths and LVT devices in the read path, which is also shown in Fig. 3.7.	49
Fig. 3.13 Simplified 8T SRAM schematic adopting boosted wordline scheme.	52
Fig. 3.14 Improvement of energy efficiency by boosting write performance. Additional energy overhead induced by the boosting voltage generation is not considered in this simulation.	52
Fig. 3.15 Comparison of normalized minimum energy consumption with write performance techniques.	53
Fig. 3.16 Improvement of minimum energy after adopting the column-interleaved scheme (Fig. 3.11(a)) and the boosted voltage scheme (Fig. 3.13). Multiplex ratio of 32 is assumed.	55
Fig. 3.17 Comparison of normalized SRAM minimum energy consumption.	55
Fig. 4.1(a) Proposed 9T SRAM cell implemented with HVT devices in write paths and LVT devices in read port. (b) Layout of the 9T SRAM cell based on 65 nm logic rules.	60
Fig. 4.2(a) 9T read SNMs compared with 6T and 10T SNMs with different voltages. (b) Distribution of 9T read SNMs at 0.4 V.	62
Fig. 4.3 Comparison of write margins of HVT and SVT devices.	63
Fig. 4.4(a) Conventional bitline sensing in the 8T SRAM [13]. (b) Concept of proposed 9T bitline sensing improvement by bitline leakage equalization technique.	64
Fig. 4.5 Improved RBL swing and sensing window of 9T bitline at 0.2 V and $f_{CLK} =$	

50 kHz with the worst case of leakage.	66
Fig. 4.6 Histogram of RBL swings of 9T SRAM at 0.2 V with the worst case of leakage from 10k-point Monte Carlo runs.	67
Fig. 4.7 Improved RBL swing with different numbers of cells and temperature. Typical corner is used in the simulation.	67
Fig. 4.8 Definition of data flipping delay and data full development delay. Difference of the data flipping delay and the full development delay substantially increases with scaling VDD.	69
Fig. 4.9 Read failure due to data non-full development in SRAM cell nodes.	70
Fig. 4.10(a) Read and write delays against scaling VDD at TT corner. (b) Read and write delays against scaling VDD at FNTP corner.	71
Fig. 4.11 Data paths of write and read operations in the CAM-assisted SRAM circuit.	72
Fig. 4.12 Delay of four different operations of SRAM and CAM circuits.	73
Fig. 4.13 Circuit diagram of CAM array, search logics and miniature SRAM array.	74
Fig. 4.14 Timing diagram of SRAM array and CAM circuit during succession of write and read operations.	75
Fig. 4.15 Faster write completion in CAM array than SRAM array at different corners.	76
Fig. 4.16 Measured (a) leakage current of the test chip and (b) write, read and average power at maximum operating frequency.	78
Fig. 4.17 Measured (a) read access time and (b) improved operating frequency of the CAM-assisted SRAM.	78
Fig. 4.18 Measured energy of SRAM only and the CAM-assisted SRAM.	79

Fig. 4.19(a) Readout waveforms capture at 0.26 V. (b) Die micro-photograph.....	80
Fig. 5.1(a) Schematic of proposed 12T dual-port SRAM cell. (b) Layout of the 12T dual-port cell.	85
Fig. 5.2(a) Leakage problem in conventional 2T read port. (b) Read bitline leakage suppression by implementation of virtual ground technique.	87
Fig. 5.3 Implementation of virtual ground technique.	88
Fig. 5.4(a) Cell stability issue in common-row-different-column access. (b) Cell stability issue in common-row-common-column access.	89
Fig. 5.5 Comparison of worst SNM scenarios in conventional DP SRAM cell and proposed DP SRAM cell.	90
Fig. 5.6(a) Illustration of read disturb in the 8T DP SRAM cell. (b) Read disturb suppression in the 12T DP SRAM cell.	91
Fig. 5.7 Simulated waveforms of read disturb for the 8T DP cell and the 12T DP cell at $V_{DD} = 0.4$ V, FN _{SP} corner. Note that the data in the 8T cell flips due to the read disturb whereas the data in the 12T cell maintains.	92
Fig. 5.8 Comparison of read SNMs of the 8T DP SRAM and the 12T DP SRAM.	93
Fig. 5.9(a) Write disturb illustration of the conventional 8T DP SRAM. (b) Write disturb suppression from the 12T DP SRAM.	94
Fig. 5.10 Circuit of hierarchical write bitline.....	95
Fig. 5.11 Architecture of the 65 nm test chip.	96
Fig. 5.12(a) Measured leakage current. (b) Measured read access time.	96
Fig. 5.13(a) Measured power consumption of read and write. (b) Measured energy per operation.....	97
Fig. 5.14(a) Micro-photograph of the test chip. (b) Captured waveforms of the RBL	

at VDD = 0.4 V.....	97
Fig. 6.1 Schematic of proposed charge-pumped DFF.....	103
Fig. 6.2 Simulated output waveforms of the two charge pumps at 0.2 V, 1 kHz. ..	104
Fig. 6.3 NMOS threshold voltage vs. transistor width at different supply voltages with a 90 nm CMOS technology [61].	106
Fig. 6.4 Simulated output waveforms of CPDFF, TGFF and ACFF at 0.4 V.....	109
Fig. 6.5 Monte Carlo simulation results of C-Q delay: (a) data ‘0’ and (b) data ‘1’. The proposed CPDFF shows less variability.....	110
Fig. 6.6(a) Setup time of CPDFF, TGFF and ACFF at different process corners and (b) Hold time of CPDFF, TGFF and ACFF at different process corners.	111
Fig. 6.7 Simulated Energy-Delay product against data activity at VDD = 0.4 V...	113
Fig. 6.8 Architecture of the FIFO circuit.....	114
Fig. 6.9 Measured C-Q delay against VDD.	114
Fig. 6.10(a) Measured power against VDD. (b) Measured power against frequency.	115
Fig. 6.11 Measured energy-delay product.....	116
Fig. 6.12 Measured power of the 2 FIFOs at 0.3 V with 10% data activity.....	117
Fig. 6.13(a) Screen capture of CPDFF output waveforms at 0.18 V and (b) die micro-photograph.	117

List of Tables

Table 3.1 Parameter summary on energy analysis simulation	42
Table 4.1 Design metric comparison with various ultra-low voltage SRAMs.....	80
Table 6.1 Performance improvement from INWE-aware sizing strategy.	108

Summary

The aggressive CMOS technology shrinking driven by cost reduction, performance improvement and power minimization enables integration of billions of transistors onto a single chip. State-of-the-art System-on-Chips (SoCs) incorporate more cores, larger capacity caches and more application-specific hardware accelerators, resulting in significant increase of power density. To reduce power and improve energy efficiency, ultra-low voltage operation is widely employed. By lowering the supply voltage from nominal level to near or beneath transistor's threshold voltage (known as near-/sub-threshold operation), the power is substantially suppressed and the energy efficiency is optimized. However, various challenging issues including high process-voltage-temperature (PVT) variation sensitivity and lack of systematic design methodology exacerbate the utility of ultra-low voltage circuits. New design methodology with minimum energy consideration to enhance performance, combat variability and suppress leakage is worthy of extensive and in-depth explorations.

In the thesis, the characteristics of transistors at near-/sub-threshold region are studied and their impact on energy consumption is investigated. Based on that, ultra-low voltage circuits with improved performance, enhanced variation-resilience and high energy/energy-delay efficiency are developed.

The main goal of the research is to explore and demonstrate optimal solutions of Static Random-Access Memory (SRAM) and D Flip-Flop (DFF) circuits in energy or energy-delay space and overcome the limitations imposed by ultra-low supply voltage. Specifically, the outcomes are demonstrated through an ultra-low voltage 9-transistor (9T) single-port SRAM, a near-threshold 12-transistor (12T) dual-port SRAM and an ultra-low voltage, energy-delay-efficient 16-transistor (16T) DFF:

- 1) As preliminary work, energy efficiency analysis of single-port SRAM utilizing multi-threshold CMOS (MTCMOS) technology is presented. The work investigates various device combinations and reveals the optimum device selection for the best energy efficiency from a MTCMOS perspective.
- 2) A 9T SRAM macro is developed with MTCMOS technology to enhance read performance and at the same time minimize leakage. In the 9T SRAM

cell, a 3T-based novel read port is proposed to equalize read bitline (RBL) leakage and improve RBL sensing margin. To optimize energy efficiency, a miniature Content-Addressable-Memory-assisted (CAM-assisted) circuit is integrated to conceal the slow data development after data flipping in write operation and therefore enhance the operating frequency. A 16 kb SRAM test chip is fabricated in 65 nm CMOS technology. The operating voltage of the test chip is scalable down to 0.26 V. Minimum energy of 2.07 pJ is achieved at 0.4 V with 40.3% improvement. Energy efficiency is enhanced by 29.4% between 0.38 V ~ 0.6 V.

- 3) A 12T dual-port SRAM is proposed to suppress disturb at the common-row-access mode and improve read-ability, write-ability and cell stability. The novel dual-port SRAM cell significantly relaxes the probability of suffering disturb, increases the resilience against disturb and extends the operating voltage to near-threshold region. In addition, hierarchical bitlines and virtual ground schemes are employed to further improve performance and leakage of the SRAM circuit. A fabricated 16 kb 12T dual-port SRAM circuit shows successful dual-port operations down to 0.4 V at the common-row-access mode.
- 4) A 16T DFF with a low energy-delay product for sub-threshold applications is presented. The device count of the proposed DFF is minimized by eliminating the clock buffer and replacing transmission gates with pass gates. To reduce the Clock-to-Q delay and improve variation resilience, two charge pumps and inverse-narrow-width-effect-aware sizing strategy are utilized, improving the performance by 23%. The fabricated DFF is fully functional down to 0.18 V and shows an energy-delay product of 13.1 pJ·ns at 100% data activity, achieving an improvement of 51.8% compared to the transmission-gate FF.

Chapter 1 Introduction

1.1 Motivation

Semiconductor process technology continues to scale by $0.7\times$ every 2 years as Fig. 1.1 depicts [1]. The aggressive CMOS technology shrinking driven by cost reduction, performance improvement and power minimization enables integration of billions of transistors into a single chip with only hundreds of mm^2 area. State-of-the-art System-on-Chips (SoCs) incorporate more cores, larger capacity caches, more radio frequency components and more application-specific hardware accelerators than ever, resulting in significant increase of power density [1]-[4]. The power consumption trend of server processors from 1990 to 2010 is revealed in [5]. The total power augments roughly 20 times over 10 years with increasing contribution from leakage power. The prevailing of big data and Gbps applications intensifies this trend and makes power minimization imperative.

Fig. 1.1 Process feature size trend [1].

Fig. 1.2 Power dissipation as a function of VDD for the 16 kb 9T SRAM [6].

The most straight-forward method to reduce power consumption is voltage scaling. By lowering the supply voltage from nominal level to near or beneath transistor's threshold voltage (known as near/sub-threshold operation), the power is substantially suppressed by several orders of magnitude. As Fig. 1.2 exhibits, the active power of the 16 kb 9-transistor (9T) Static Random-Access Memory (SRAM) decreases from 1.3 mW to 1.2 μ W when supply voltage changes from 1.2 V to 0.24 V with power saving of more than 1000 times [6]. Due to the attractive prospective, substantial research activities about ultra-low voltage circuit have been performed [7]-[10]. However, various challenging issues including frailty sub-threshold operations, high process-voltage-temperature (PVT) variation sensitivity and lack of systematic design methodology exacerbate the utility of ultra-low voltage circuits. New design methodology with novel circuit techniques to enhance performance, combat variability and suppress leakage is worthy of extensive and in-depth explorations.

Ultra-low voltage operation improves not only power dissipation but energy efficiency as well. The emergency of energy constrained applications, such as

Fig. 1.3 Energy dissipation as a function of VDD for the 16b 1024-point FFT [7].

handheld devices, wearable electronics, wireless sensor nodes, and implantable biomedical instruments demand energy efficient circuit solutions to prolong battery life. Although energy harvesting circuits are widely used in energy autonomous systems, the energy it derives from the external sources is a very small amount and cannot satisfy the energy specification of the whole SoC. Now it is known that energy has a correlation with supply voltage. As Fig. 1.3 presents, energy per operation of the 16b 1024-point Fast Fourier Transformation (FFT) decreases with voltage scaling. However, the energy as a function of supply voltage reaches an optimal point and increases again even if the voltage is lowered further. Thereby, how to find the optimal point in the energy space to ensure energy efficient circuit is of high importance.

While energy is paramount for circuits like the Random Access Memory (RAM), performance is equally critical for various other circuitries such as the standard cell logics with respect to the traditional design philosophy. For these circuits, e.g. flip-flops (FFs), delay is *as crucial as* energy [11][12]. To emphasize both parameters, extended specification space has to be adopted to evaluate the performance-sensitive circuits. Hereby, energy-delay product is exploited as a key metric to evaluate them.

1.2 Research Objectives and Contributions

Numerous researches have fueled the field of ultra-low voltage circuits but the subsequent challenges, such as variability, leakage and etc., become more and more severe as transistor length shrinks below 100 nm. In the thesis, we study the characteristics of transistors at near-/sub-threshold region and investigate its impact on energy and energy-delay product. Based on that, we aim to develop ultra-low voltage circuits with improved performance, enhanced variation-resilience and high energy/energy-delay efficiency.

The research outcomes are demonstrated through ultra-low voltage SRAMs, D flip-flop (DFF) and First-In-First-Out (FIFO) circuits. Specifically, the contributions of our research work are as followings:

1. Minimum-energy-driven SRAM design is highly sought after in numerous emerging applications. As preliminary research, the thesis presents SRAM energy analysis utilizing multi-threshold voltage (multi- V_{th}) devices and various circuit techniques for power reduction and performance improvement, and suggests optimal device combinations for energy efficiency improvement. In general, higher- V_{th} devices are preferred in the cross-coupled latches and the write access transistors for reducing leakage current while lower- V_{th} devices are desired in the read port for implementing higher performance. However, excessively raised V_{th} in the write paths, i.e. the cross-coupled latches and the write access transistors, leads to slower write speed than read, which quickly nullifies improved energy efficiency. In the work, the energy efficiency improvement of 6.24 $\times$ is achieved only through an optimal device combination in a commercial 65 nm CMOS technology. Employing power reduction and performance boosting techniques together with the optimal device combination enhances the energy efficiency by up to 33 $\times$.
2. Conventional 6-transistor (6T) SRAMs suffer severe cell stability issue during read at ultra-low voltage. Decoupled SRAM cells, such as 8T SRAM cell [13], are widely adopted to ameliorate this issue by decoupling read port

from the data storage latch. Based on this preliminary research, a 9-transistor (9T) SRAM cell is developed using the multi-threshold CMOS (MTCMOS) technology to enhance read current and speed while minimize the leakage current. Another issue of the 6T SRAMs at ultra-low voltage is the degraded read sense ability due to data-dependent leakage. In the 9T SRAM cell, a 3T-based novel read port is proposed to equalize read bitline (RBL) leakage and to improve the RBL sensing margin by eliminating the data-dependence of bitline leakage current. To optimize energy efficiency, a miniature CAM-assisted circuit is integrated to conceal the slow data development after data flipping in write operation and therefore enhance the operating frequency. A 16 kb SRAM test chip has been fabricated in 65 nm CMOS technology. The operating voltage of the test chip is scalable from 1.2 V down to 0.26 V with the read access time from 6 ns to 0.85 μ s. Minimum energy of 2.07 pJ is achieved at 0.4 V with 40.3% improvement compared to the SRAM without the aid of the CAM. Energy efficiency is enhanced by 29.4% between 0.38 V ~ 0.6 V by the proposed CAM-assisted circuit.

3. Dual-port SRAMs can execute two operations simultaneously in one clock cycle. Current 8T dual-port SRAMs are implemented in a similar way as the conventional 6T SRAMs. Apart from the weakness inherited from the 6T SRAMs, the 8T dual-port SRAMs are challenged by common-row-access disturb, which severely limits their operation under low voltage. In the thesis, a 12-transistor (12T) dual-port SRAM is proposed to suppress the disturb at the common-row-access mode and improve the worst case read-ability, write-ability and cell stability. The work significantly relaxes the probability of suffering disturb, increases the resilience against disturb and extends the operating voltage to near-threshold region. In addition, a virtual ground technique is employed to further lower the power and energy by reducing bitline leakage. A 16 kb 12T dual-port SRAM has been fabricated in a 65 nm CMOS process technology and showed successful dual-port SRAM operations down to 0.4 V at the common-row-access mode.
4. Analysis of energy-delay domain reveals that performance boosting

technique is essential to achieve an optimal energy-delay product. This is exactly what the research on ultra-low voltage DFF achieves. A 16-transistor DFF featuring a low energy-delay product for near-/sub-threshold applications is implemented. The device count of the proposed DFF is less than the mainstream DFF, such as the transmission-gate FF (TGFF). This is possible through eliminating clock buffer and employing pass gates instead of transmission gates. To reduce the Clock-to-Q (CQ) delay and improve its variation resilience, two charge pumps and an inverse-narrow-width-effect-aware strategy are utilized, improving the performance by 23%. The novel DFF fabricated with 180 nm CMOS technology is fully functional down to 0.18 V and shows an energy-delay product of 13.1 pJ·ns at 100% data activity, achieving 51.8% improvement compared to the conventional TGFF, respectively. When $V_{DD} = 0.5$ V, the energy-delay product is averagely enhanced by 50.8%. Two 256-bit FIFOs are implemented using the proposed DFF and TGFF. The FIFO utilizing the charge-pumped DFF exhibits 31.2% total power reduction at subthreshold regime.

In summary, the main contribution of the thesis is exploration and demonstration of optimal solutions for SRAM and DFF circuits which overcome the limitations by ultra-low voltage while satisfy the requirements of energy constrained applications.

1.3 Organizations

The rest of the thesis is organized as follows. Chapter 2 introduces the background for ultra-low voltage, low power SRAM and flip-flop circuitries. Common design techniques to assist near-/sub-threshold operation and minimize PVT variation are reviewed. Fundamental knowledge of energy/energy-delay efficient design methodology is provided. Chapter 3 models the energy consumption of an 8T SRAM and comprehensively investigates the impact of multi-threshold CMOS (MTCMOS) devices on energy efficiency. Subsequently, various assisting-circuit techniques for power reduction and performance improvement are examined. Optimal device combinations for energy efficiency improvement are suggested. Chapter 4 presents an in-depth analysis on the correlation of MTCMOS technology on energy efficiency. Based on the observation, the methodology to design an energy-efficient MTCMOS SRAM with improved read sensing margin and enhanced write performance is discussed. Specifically, the RBL leakage is analyzed and leakage equalization is utilized to make a higher read sensing margin. Read and write delays are investigated and the idea of exploiting a miniature Content-Addressable-Memory (CAM) circuit is investigated to boost the write performance and ultimately the energy efficiency. Chapter 5 analyzes the common-row-access behavior for the conventional 8T dual-port SRAM. The solution to suppress disturb at the common-row-access mode is explored by a novel 12T dual-port SRAM. Other techniques, such as virtual ground and hierarchical bitline are also evaluated in the chapter. The DFF work is presented in Chapter 6. In this chapter, the Clock-to-Q delay parameter is probed and optimized through the sizing methodology and use of charge-pump circuits. Silicon measurement results are included to validate the effectiveness of the techniques on energy-delay improvement. Chapter 7 summarizes the entire research work in the thesis and looks ahead the possible future works.

Chapter 2 Background and Literature Review

2.1 Conventional 6T Single-port SRAMs

Single-port Static Random-Access Memory (SRAM) has been widely utilized in CPUs and processor cores as on-chip memory to provide solution to intermediate data access. Compared to Dynamic Random Access Memory (DRAM), SRAM enables static on-chip data storage. This feature makes SRAM less complicate than DRAM which has to be refreshed periodically in order to retain data. Without the additional circuitry and timing to introduce the refresh, SRAM is generally faster and less power hungry than DRAM. In CPUs, the embedded SRAMs usually serve as cache memories due to its speed, density and energy characteristics.

2.1.1 6T Single-port SRAM Operation

The conventional 6T single-port SRAM is depicted in Fig. 2.1. Transistors $M1$ and $M2$ are switches for data access. Transistors $M3\sim M6$ form a cross-coupled latch in the middle and serve as a data storage element. The node Q and QB hold data and its opposite value. Column-wise, hundreds of SRAM cells are assembled and share the data in and out paths, which are known as bitline (BL) and bitline bar (BLB). The pair of bitlines is connected to the drain terminals of all the access transistors in

Fig. 2.1 Schematic of conventional 6T SRAM cell.

Fig. 2.2(a) Concept of read operation in a 6T SRAM bitline. (b) Concept of write operation in a 6T SRAM bitline.

the column. Row-wise, dozens of SRAM cells are connected to each other by a shared wordline (WL), which are utilized to activate operations.

Fig. 2.2(a) and (b) illustrate read and write operations of the 6T SRAM cell, respectively. At low level of a clock cycle, the bitlines (BL and BLB) are pre-charged to VDD or a moderate high voltage. In read operation, WL is asserted to turn on the access transistors while the preliminarily pre-charged bitline pair is left floating for read evaluation. According to the data pattern, BL and BLB can behave differently. When Q holds logic '1', M3 and M6 are turned on while M4 and M5 are cut off. Hereby, BLB can discharge mainly through a path formed by M2 and M6. If, on the contrary, if Q holds logic '0', M3 and M6 are switch off whereas M4 and M5

are turned on. The conductive path forms from $M1$ to $M5$, which decreases the voltage of BL. At the end of the bitline pair, a sense-amplifier responds to the voltage difference between BL and BLB to output the value of the cell as long as the cell established an enough voltage difference in the cycle.

In write operation, data and its opposite value are loaded into BL and BLB, respectively. Specifically, one bitline is pre-charged to a high voltage and the other is connected to ground. The access transistor $M1$ and $M2$ are simultaneously switched on by raising the WL voltage to VDD. Since the strength of the NMOS pass-gate is sufficiently stronger than the PMOS pull-up device, the internal node with logic '1' is pulled down by the adjacent bitline which is grounded. The positive feedback due to the cross-coupled structure assists to flip the original value and maintain the new data.

2.1.2 Challenges of 6T SRAMs for Ultra-low Voltage Operation

Although the 6T SRAM circuit is mature in industry and used in most commercial chips, it is very poor in voltage scalability. The reasons can be categorized into three aspects. Firstly, the read disturb and the write-ability issue impedes voltage scaling. Secondly, the large variation at ultra-low voltage can cause severe reliability problem. Lastly, degraded I_{on} -to- I_{off} ratio makes sub-threshold operation very difficult.

The read operation of 6T SRAMs is fast but destructive, that is, the cell nodes can suffer disturbance during read operation and the data can be overwritten by disturb current. Fig. 2.3 depicts the read stability issue. The data storage node Q and QB are directly accessed to the bitline pair through the access devices. Therefore, due to voltage division effect between the cross-coupled latch and the bitline capacitance, the value in the SRAM cell is vulnerable to flip during read. Specifically, voltage at node QB rises to a small amount ΔV above ground when $M2$ discharges BLB. If ΔV is larger than the trip point of the inverters, cell value is flipped with the effect of the positive feedback loop. To prevent the destructive read, access transistors $M1$ and $M2$ are required to be downsized or the pull-down transistors $M5$ and $M6$ are needed to be upsized. In other words, the cell ratio which is defined as the ratio of drain current of the pull down device and the drain current of the access device has to be increased accordingly to suppress ΔV . To evaluate the data stability of the 6T

Fig. 2.3 Cell stability degradation of 6T SRAM cell due to read disturb.

Fig. 2.4 Illustration of SNM of 6T SRAM cell.

SRAM cell, read Static Noise Margin (SNM) is adopted as a key functionality metric for read operation. It is defined as the minimum amount of DC noise required to flip the state of the cell. Fig. 2.4 illustrates the SNM in the DC transfer function curves of the latch. Normally, the larger the SNM is, the better the ability of the cell against read disturb is. But read SNM deteriorates significantly with voltage scaling.

On the other hand, the bitlines have to overpower cell with new data to foster a successful write operation. During write cycle, the grounded bitline provides a discharging path for the internal node holding logic '1'. The relative strength of the access transistor and the pull-up PMOS device determines the write-ability. To ease data flipping, the pull-up strength should be smaller than the access strength to weaken data retention capability. Usually, write margin is employed as a key metric to evaluate the write-ability of SRAM cells. For 6T SRAMs, it is interpreted as the voltage headroom at WL for a successful write operation.

As [14],[15] manifest, the minimum operating voltage is bounded by read stability and write margin. However, the substantial degraded read SNM and write margin of 6T SRAMs prevent aggressive voltage scaling and make sub-threshold operation extremely challenging without the aid of assisting circuits. Fig. 2.5(a) and (b) exhibit the trends of read SNM and write margin with scaled VDD, respectively.

The SNM is far less than half VDD at all supply voltages while the write margin degrades approximately 9× when the voltage scales from 1.2 V to 0.2 V.

The degradation of the two key metrics is accelerated by large V_{th} variation at ultra-low voltage. Conventionally, transistor is thought to cutoff when the overdrive voltage ($V_{GS}-V_{th}$) becomes zero. In fact, transistor is still working although the V_{GS} is lower than the threshold voltage V_{th} and the current follows a correlation defined by Equation (2.1):

$$I_{sub} = I_0 \cdot \frac{W}{L} \cdot e^{q(V_{GS}-V_{th})/(n \cdot kT)} \quad (2.1)$$

where I_0 is highly relevant to the technology, q is the electronic charge, T is the temperature and k is the Boltzmann's constant. For sub-threshold computing, since the current is exponentially correlated to $(V_{GS}-V_{th})$ and temperature, very small drifting of VDD, V_{th} and temperature can cause large amount of current variation. Therefore, the PVT (process, voltage and temperature) variation which is mainly attributed to voltage scaling, random dopant fluctuation and temperature variation worsens various figures-of-merit and creates reliability issue, such as unacceptable bit error rate. To combat variability, supply voltage of 6T SRAM circuits cannot be tuned to very low.

Leakage is another bottleneck to overcome for ultra-low voltage operation. When VDD is high, the drain current of transistor is tens of thousands times larger than the leakage current. The high I_{on} -to- I_{off} ratio at strong inversion region makes the impact of leakage current on key design metrics negligible. However, if circuits work at near- or sub-threshold region, the transistor channel is only moderately or weakly inverted, resulting in a small I_{on} -to- I_{off} ratio. It is worsen by CMOS technology shrinking which makes gate leakage more and more difficult to control.

When the leakage current is comparable to the drain current, characteristics of 6T SRAM circuits, such as operating frequency, read sensing ability, energy efficiency change accordingly. In addition, bitline leakage dependent upon data pattern of each column in 6T SRAMs is detrimental for read sensing. If the amount of the data-

(a)

(b)

Fig. 2.5(a) 6T SRAM read SNM by VDD sweeping. (b) 6T SRAM write margin by VDD sweeping.

dependent bitline leakage is considerable enough, the bitline level of data '1' could be lower than that of data '0' [16]. Consequently, this limits the number of cells per bitline and the minimum operating voltage.

2.1.3 Design Techniques of 6T SRAMs to Improve Minimum Voltage

Diverse design techniques are proposed to cope with the challenges and improve the minimum operating voltage of 6T SRAMs. Basically, most of these techniques aims to improve read SNM, enhance write-ability and reduce leakage.

Decoupled SRAM cells with dedicated read ports such as 8T and 10T SRAM cells [13],[17],[18] are common solutions to suppress read disturb and improve cell stability. The read port, which is decoupled from the internal storage node by separation from the data storage element, enables a single-ended read sensing without internal node access. When read is asserted, the read current flows through the transistors in read port without any interference to Q and QB. As the read disturb is diminished, smaller transistor can be adopted in SRAM cells to compensate the area overhead. In [13], a two-transistor read stack is added to the standard 6T cell and enabled by read wordline (RWL) as Fig. 2.6(a) shows. As RWL is asserted, *M7* is on and the exclusive read bitline (RBL), which is precharged in prior, is left floating for data evaluation. When Q stores logic '1', *M8* is cutoff due to negative overdrive voltage. Therefore no conductive path forms in the read port, which maintains a high voltage of RBL. When Q stores logic '0', *M8* is switched on and the read current forms from RBL through *M7* and *M8* to ground. Consequently, the RBL voltage is fast pulled down and amplified for sensing data '0'. The 8T SRAM, combined with optimized write pass-gate and MTCMOS technology, achieves 295 MHz operation at 0.41 V. Alternative technique, such as wordline underdrive, is also effective to minimize read disturb by driving WL lower than VDD. Compared to the decoupled SRAM cells, it incurs less area penalty which is beneficial for high-density SRAMs [19].

Write margin is another challenge to overcome. The sizes of the pull-up PMOS transistors and the NMOS access transistors have to be carefully designed to ensure the write current is strong enough to flip the original data. Moreover, the strength of the write access transistors is susceptible to PVT variations, which is severe in advanced technologies. If the access devices are weakened by the PVT variations, the write margin could be degraded. To assist write and improve write margin,

(a)

(b)

Fig. 2.6(a) Schematic of 8T SRAM cell [13]. (b) Schematic of 10T SRAM cell [17].

various design techniques are utilized to manipulate wordline, bitline and cell supply voltage. Boosted wordline technique pumps the voltage of WL higher than VDD to enhance V_{GS} and ease data flipping during write. Collapsed cell supply voltage method intentionally decreases the supply voltage of the cross-coupled latch during write [17] to weaken the hold SNM. Negative bitline strengthens the access transistor by pull down the voltage of the bitline with data '0' to a negative voltage which enforces write margin ultimately [19],[20],[21]. In [21], a negative bitline is implemented by using a negative bias to represent data '0'. This write-assist technique enhances the write ability through increasing the strength of the access

transistor in the SRAM cell. However, it requires a tracking replica bitline, a pulse generator and a negative charge pump, which incurs increased control complexity and additional circuitry.

For bitline with a large number of cells, leakage current becomes more significant and causes bitline sensing more difficult with voltage scaling. The 10T SRAM cell [17] depicted in Fig. 2.6(b) exhibits a read port for leakage reduction. The source of $M3$ and $M4$ are connected to a cell supply V_{DD} for write. The read port is implemented by $M7$ through $M10$. As Q holds '0' and QB holds '1', $M10$ adds an off device in series with the leakage path through $M8$ and the path through $M9$. As Q holds '1' and QB holds '0', $M10$ reduces leakage through $M7$ by the stack effect. The reduction in sub-threshold leakage through $M8$ reduces the impact of leakage from unaccessed cells and allows more cells on a bitline. The 256 kb 65 nm CMOS test chip using the 10T cell and boosted wordline scheme functions without error at 380 mV. At 27°C, the test chip approximately dissipates 2 μ W in terms of leakage power with a supply voltage of 0.3 V. The 10T memory saves over 60 $\times$ in leakage power when V_{DD} scales from 1.2 V to 0.3 V [17].

To address the data-dependent bitline leakage challenge, circuit techniques are explored and proposed. In [22], the pull-down bitline leakage in the conventional 6T SRAMs is compensated by injecting additional pull-up current. However, the analog detection and injection circuit in this design is highly sensitive to process variations and coupling noises [23]. Sensing calibration technique proposed in [24] solves the data-dependency problem by injecting the same voltage offset on BL and BLB using a crossing-structure circuit. The main drawback is the calibration scheme, which needs bitline loading capacitors and hence the higher power consumption. Although the P-P-N based 10T SRAM cell presented in [25] achieves high immunity to the data-dependent bitline leakage, it has to triple the size of the four transistors to improve its write ability and thus occupies substantial area.

2.2 Conventional 8T Dual-port SRAMs

Dual-port SRAMs can read and write different cells at different addresses simultaneously. This increases bandwidth by approximate $2\times$ compared to single port SRAMs, which accesses only a cell at a time. As design complexity grows, greater demands are placed upon high-bandwidth memories to boost throughput. For example, high-speed communication and multi-media processing [26]-[29] need dual-port SRAMs to improve the total chip performance by parallel operation. In addition, dual-port SRAMs can be implemented as register file in CPU. There are two types of dual-port SRAMs, which are synchronous dual-port SRAMs and asynchronous dual-port SRAMs. In our research work, only synchronous dual-port

(a)

(b)

Fig. 2.7(a) Schematic of conventional 8T dual-port SRAM. (b)

Parallel memory access of 8T dual-port SRAM [30].

SRAMs are investigated.

2.2.1 8T Dual-port SRAM Operation

Fig. 2.7(a) shows the schematic of the conventional 8T dual-port SRAM cell. Port A and Port B are accessed by exclusive address and operation instruction. Each port consists of their corresponding wordline (WL) and a pair of bitlines (BL and /BL).

Fig. 2.8(a) Illustration of different-row-different-column access. (b) Illustration of different-row-same-column access. (c) Illustration of common-row-different-column access. (d) Illustration of common-row-common-column access [30].

Combined with the cross-coupled latch, the operation of each port acts exactly like a 6T single-port SRAM cell, which has been analyzed in details in Section 2.1.1. The widths of the two NMOS drive transistors are expanded to maintain the cell stability against common-row-access. Parallel memory access by the dual-port SRAM block is portrayed in Fig. 2.7(b), where both unit-A and unit-B can access a dual-port SRAM cell simultaneously within a cycle. Accordingly, there are four possible simultaneous access operations occurring at the same address: read-write, write-write, read-read, and write-read. A simultaneous read-read operation does not affect the cell but the other operations need proper measures to ensure that no data collision occurs. Fig. 2.8 presents the variety of the access situations of the dual-port SRAMs when both ports are activated within a clock cycle. Fig. 2.8(a) shows the case in which a SRAM cell is accessed from both ports designated independently by each address. Fig. 2.8(b) describes a situation of different rows but a common-column-access. Both of the two cases have no access conflict issue because either of the SRAM cell is only enabled by one port, which is exactly like a single-port SRAM operation. Fig. 2.8(c) and (d) present situations incurring access conflicts, which are common-row-different-column access and common-row-common-column access, respectively. In these common-row-access cases, port A and port B of the selected row are simultaneously enabled, which substantially degrades SNM and write margin, which will be elaborated in Section 2.2.2.

2.2.2 Challenges of 8T SRAMs for Ultra-low Voltage Operation

Conventional 8T dual-port SRAM cell is derived from the standard 6T single-port SRAM cell. It inherits the weakness of the 6T SRAM cell at ultra-low voltage like poor cell stability, reduced read-ability and write-ability, which impedes voltage scaling. The above issues are exacerbated because the 8T dual-port cell has two more access transistors exposure to disturbance, especially in the common-row-access mode. This limits the minimum operating voltage which is generally not as low as that of the 6T single-port SRAM circuit.

In the common-row-access cases, the cell stability has to be treated as the worst case because the enabled two WLs impose more disturbances and degrade SNM for all cells along with the selected row. Fig. 2.9 exhibits the SNM in common-row access and different-row access [30]. The butterfly curve of the 8T dual-port SRAM cell is depicted by overlapping the voltage transfer curve of one inverter with its inverse. The SNM is visualized by the diagonal length of the largest square which can be embedded between the two lobes of the butterfly curve. For common-row access mode, the read SNM is reduced approximately by 1/3, which directly results from the degradation of electrical β ratio. The β ratio is deteriorated from β_{D1}/β_{A1} (different-row access) to $\beta_{D1}/(\beta_{A1} + \beta_{A2})$ (common-row access) because of the simultaneous activation of WLA and WLB, where β_{D1} , β_{A1} , β_{A2} indicate the coefficients of source-drain currents of the pull-down NMOS transistor, the access transistor for port A and the access transistor for port B.

In 8T dual-port SRAM bitline, write-ability has to be also taken account of as the worst case. When port A writes, the cell incurs disturb due to the activation of the other port, which prevents the data flipping by current interference with the storage node. The situations of the read and the write disturb will be analyzed in Chapter 5.

As discussed, the situations in Fig. 2.7(c) generate the worst case of read and write for a selected cell and the worst case of data retention for unselected cells. Same address access for write operation, as shown in Fig. 2.7(d) is prohibited because the destructive write causes abnormal leakage current in the SRAM cell if the writing

Fig. 2.9 Comparison of read SNMs in different-row access and common-row access situations [30].

data from both ports are opposite [30]. Still, the simultaneous write-read, read-write and read-read operations are allowed and frequently required from the system. As the conventional dual-port SRAM must satisfy various worst case situations, the size of the memory cell has to be increased accordingly to improve cell stability as well as read and write abilities. Normally, the drive NMOS transistor in the 8T dual-port cell can be oversized more than $2\times$ as the transistor size in the 6T single-port cell.

In summary, the conventional 8T dual-port SRAM is extremely poor in terms of voltage scalability due to its intrinsic flaw of cell stability and the difficulty to perform read and write operations at ultra-low voltage.

2.2.3 Design Techniques of 8T SRAMs with Ultra-low Supply Voltage

Most of the contemporary design techniques for single-port 6T SRAMs are universal for SRAM circuits. Hereby many of them are transplanted to dual-port SRAM to enhance figures-of-merit. Recent state-of-the-art dual-port 8T SRAMs are improved in various ways, such as low leakage and conflict-free 8T solutions. Although those circuit techniques are beneficial to lower the supply voltage, very rare cases of 8T dual-port SRAM are demonstrated through ultra-low voltage environment.

A 20 nm high-density dual-port SRAM with wordline-voltage-adjustment technique has been demonstrated in [31] to achieve better read-ability and write-ability against local variation. This scheme adopts lowering wordline voltage for read assist and raising wordline voltage for write assist. A temperature-monitoring device is embedded to sense the temperature variation for more accurate control. An assisting controller is connected to the fuse which triggers the on-chip regulator to generate corresponding wordline voltage according to different process variations and temperature thresholds. Measurement reveals a 0.1 V minimum voltage improvement.

Fig. 2.10 Concept of access circumvention scheme for dual-port 8T SRAM [30].

Fig. 2.11 Concept of active bitline equalizing technique for dual-port 8T SRAM [32].

The minimum operating voltage of 8T dual-port SRAMs, as discussed in Section 2.2.2, is constrained by the situation where the same row address is asserted by both ports simultaneously. The read disturb and write disturb together with the cell stability challenge under the circumstance substantially impede voltage scaling. Therefore, how to eliminate the conflict in common-row access becomes the key to lowering supply voltage further and draw great attentions for researchers. A circumvention technique has been proposed to avoid simultaneous common-row access for 8T dual-port SRAM [30]. It adopts a priority row decoder and a bitline shifter to prevent access conflict. The fundamental concept of the scheme is illustrated in Fig. 2.10. In the scheme, port A is defined as primary whereas port B is considered as secondary. A row address comparator and a bitline shifter are introduced in the secondary port. When both ports are asserted at the same row, the comparator outputs logic '0' to disable the secondary row decoder for port B. Consequently, only the wordline for port A is accessible to the SRAM cell. Meanwhile, the logic '0' from the comparator changes the connection of the secondary port from the pair of BLB to that of BLA to foster a possible read operation. However, this technique incurs area penalty because of the complexity of the comparator and the bitline shifter circuits. More importantly, the circumvention

Fig. 2.12(a) Write-disturb detector for 8T dual-port SRAM. (b) Coordinately-activated write drivers for dual-port SRAM [32].

scheme delays the operation through the secondary port in common-row access mode, which degrades the performance of the dual-port SRAM. The measurement results shows the 32 kB macro fabricated with 65 nm technology can work down to 0.8 V.

An active bitline equalizing circuitry has been proposed to improve the write margin by removing write disturb when the same row address is activated [32]. Fig. 2.11 presents the concept of the equalizing technique. In common-row-access mode,

the pair of bitlines in port B (BLB) is disconnected from the storage node and equalized to the pair of bitlines in port A (BLA), thus the write disturb from BLB is circumvented. To realize the concept, write-disturb detector and coordinately-activated write driver are proposed and implemented in the selected column (Fig. 2.12(a), (b)). By utilizing this circuitry, the minimum voltage of the 28 nm dual-port SRAM macro is improved by 120 mV to 0.66 V for slow clock cycle at expense of 5.8 % extra area.

In summary, although various assisting circuitries have been proposed to reinforce low voltage operation, the minimum voltage of 8T dual-port SRAMs are still relatively higher. Lower operating voltage for dual-port SRAM should be continuously pursued.

2.3 Conventional D Flip-Flop Circuits

D Flip-Flops (DFFs) are very fundamental components in digital CMOS integrated circuit (IC) design. They can account for substantial random-logic power and random area. Most DFFs are consisted of pairs of latches which are transparent on different phases of a clock cycle. In general, the latches are build upon regenerative storage type, which are “static” because data are constantly restored by the positive feedback in the storage loop. This method prevents the stored data from being corrupted by parasitic leakage current compared to capacitor storage type [33].

2.3.1 Mainstream DFF Circuits and Timing Properties

Edge-triggered FFs are very popular in globally clocked systems mainly due to the simplicity of referencing all events and timing parameters to a single toggling of the clock [33].

One of the most widely used edge-triggered FFs is transmission-gate FF (TGFF), which is composed of two level-sensitive latches but operates in two opposite clock phases controlled by corresponding transmission gate. Fig. 2.13 shows the circuit of TGFF. At the low phase of the clock cycle, the inverted input D is sampled by the first latch, which is known as the master stage, with a connected path linked by the first transmission gate. The new data overpowers the original data stored in the master latch with cutoff of the feedback loop as the transmission gate in the loop turns off simultaneously. At the high phase of the clock cycle, the second latch which is the slave stage samples the data held in the master stage and outputs it with

Fig. 2.13 Schematic of transmission-gate FF (TGFF).

Fig. 2.14 Schematic of True-Single-Phase-Clocked (TSPC) FF.

inversion. It is noticeable that the clock fed to each transmission gate is buffered locally to accommodate the high load of the two clock phases.

True Single-Phase-Clocked (TSPC) FF is a common dynamic DFF variety which cascades the negative and positive dynamic latches of Yan and Svensson. Fig. 2.14 depicts the schematic of TSPC. Transistor $M1 \sim M3$ consist of the negative dynamic latch which is transparent when the voltage of clock is low. Similarly, transistor $M7 \sim M9$ consists of the positive dynamic latch, which is transparent when the voltage of the clock is high. At the low phase of the clock signal, $net1$ samples the inverted input while $M4$ isolates the second latch from the first one. As the clock signal toggles, $M4$ shuts off and $M6$ is switched on. The voltage of $net2$ varies on the basis of the voltage of $net1$. If $net1$ stores logic '1', the voltage of $net2$ discharges through the conductive path of $M5$ and $M6$ and triggers $M7$ to raise the voltage of QB. If, on the contrary, $net1$ stores logic '0', the voltage of $net2$ will remain high to turn on $M9$. The voltage of QB discharges to ground and an output of logic '1' forms at the Q terminal.

Timing properties are important for synchronous sequential logic circuits. Any sequential logic has to comply with certain timing specifications to ensure a successful operation. For edge-triggered FFs, *set-up time*, *hold time* and *Clock-to-Q (C-Q) delay* are the most important timing metrics. Fig. 2.15 illustrates the three parameters and their concepts. The *set-up time* is defined as the delay from the data's becoming valid to the rising edge of the clock. Likewise, the *hold time* is the delay from the clock to the data's becoming invalid [34]. The *C-Q delay* is the delay

Fig. 2.15 Illustrating DFF setup time and hold time [34].

from the rising edge of the clock to the output's becoming valid. Violation of *set-up time* and *hold time* can cause *C-Q delay* to increase or even output flipping. A more practical way to measure *set-up time/hold time* is to capture the *set-up/hold skew* when nominal delay is degraded by 10% [35]. The energy-delay DFF circuit proposed by the thesis adopts this method to measure both parameters.

2.3.2 Design Challenges of DFFs for Energy Efficient Applications

The conventional TGFF utilizes two opposite level-sensitive latches to enable the edge trigger operation. It employs two inverters to build a clock buffer locally so as to increase loading capability and reduce switching current. However, the clock buffer consumes power as long as the clock toggles, even at low data activity when input D rarely changes. This persistent power consumption causes extra power and energy, which is disadvantage for energy-constrained applications when a large number of TGFFs is integrated. On the other hand, the TGFF is prone to incur timing property when it is working at near- or sub-threshold regimes. As the clock is buffered and the data is inverted, the mismatch between the clock path delay and the data path delay can cause problems. Specifically, the NMOS device in the first transmission gate can turn off earlier than its PMOS. Likewise, the PMOS in the first feedback loop turns on before its NMOS counterpart. This can cause hold time violation when the input data changes from logic '1' to logic '0' just after the clock edge [36]. Moreover, the hold time degrades at ultra-low voltage where the PVT is accentuated.

The TSPC eliminates the clock buffer by utilizing one clock phase. However, dynamic operation of the circuit degrades its robustness especially at ultra-low voltage because *net1* and *net2* are extremely subject to leakage and noise when they are not being driven. For example, when the input is logic '0' and the clock signal toggles to high level, the voltage of *net1* is neither pulled-up to VDD nor pulled-down to ground. Likewise, the voltage of *net2* does not necessarily remain high when *M4* is shut off because *M5* can be non-ideally switched off. To make things worse, once the voltage of *net2* drops to ground while the clock is high, the node would remain low because there is nothing to pull it up, resulting a functionality problem.

2.3.3 Design Techniques for Energy Efficient DFFs

Fig. 2.16 Schematic of adaptive-coupling FF (ACFF) [37].

Although the conventional TGFF circuit is robust in ultra-low voltage, it is not energy efficient due to the persistent power spent on the clock buffer in every clock cycle. On the other hand, TGFF consumes 24 devices, which can incur large random area penalty as well as leakage energy. To solve the problem, an emerging DFF utilizing adaptive-coupling configuration to reduce transistor count and power for energy saving has been proposed recently [37]. Fig. 2.16 illustrates its schematic and configuration. To remove the clock buffer, a differential master-slave topology is proposed in the adaptive-coupling FF (ACFF). The original transmission gates in the propagation path are replaced by PMOS and NMOS pass gates, respectively. However, the circuit is subject to process variations, because the PMOS pass gates are too weak to provide a strong source-drain current at low voltage to overpower the strong coupling from the latch during a transition. To ameliorate it, the adaptive-coupling method is introduced such that the strong coupling in the feedback loop is weakened when the input is opposite to the internal storage. Specifically, an adaptive-coupling element which is comprised of a PMOS transistor and an NMOS transistor is configured in parallel to control the cross-coupled loop. If the level of node *BN* is high, the PMOS device is cutoff and the NMOS device is switched on, weakening the feedback path of *G-F* and enabling easier discharging of node *F* to

node *B*. As node *FN* is charged to high state, the level of node *G* accordingly becomes low and enforces node *F* to discharge completely by the NMOS device in the adaptive-coupling element. Consequently, the C-Q delay is improved due to the easier data transition and the energy is optimized by the elimination of power-hungry clock buffer and the reduction of device count. Silicon experiments on 40 nm CMOS technology validates this scheme achieves a less mean C-Q delay with a smaller standard deviation and a reduced hold time compared to TGFF. The energy per cycle is improved by 60.8% at 10% data activity with a supply voltage of 1.1 V. Up to 77% energy reduction is obtained at 0% data activity mainly due to the elimination of the clock buffer. Despite the improvements, the ACFF circuit is reported to work typically at super-threshold region ($V_{DD} > 0.75$ V), which is not fully qualified for ultra-low voltage operation.

The dynamic TSPC circuit, as analyzed in Section 2.3.2, is highly susceptible to noise and PVT variations, thus it is very fragile for ultra-low voltage operation. An improved DFF on the basis of TSPC has been proposed to foster a static operation with single-phase clocking and contention-free transitions [38]. The so-called Static Single-Phase Contention-Free FF (S^2CFF) (Fig. 2.17) has the same transistor count as a TGFF. When *D* is '0', *net1* stores the opposite value of input data and *net2* is precharged at the low phase of the clock. As the clock toggles, *net2* discharges to ground through *M9* and *M10* and updates *QN* by switching on *M13*. When *D* is '1', *net1* discharges to ground and *net2* is pulled up to V_{DD} . At the high phase of the clock, *QN* is updated by the pull-down operation through *M14* ~ *M16*. In the S^2CFF circuit, *net1* and *net2* become static nodes, which is different from the TSPC design. This is accomplished by the clocked devices and the positive feedback loop between the two nets. As such, the operation of the DFF is fully static. On the other hand, the sub-circuit consisted of *M11*, *M12* and *M15* prevents possible glitch which enables a contention-free operation. The S^2CFF has been implemented in a 45 nm SOI technology and showed a clock power reduction of 41% and a total sequential power reduction of 39% at 1V/1GHz compared to TGFF. Active energy is also improved by 32% and 34% at $V_{DD} = 1$ V and 0.4 V, respectively. The reported minimum supply voltage of the S^2CFF is 0.4 V.

Fig. 2.17 Schematic of Static-Single-Phase Contention-Free FF (S^2CFF) and its operation [38].

In summary, the ACFF and the S^2CFF exhibit substantial improvement on power and energy efficiency compared to TGFF. However, the minimum supply voltage of each DFF is just moderate. Further exploration on ultra-low voltage and minimum energy-driven techniques which are effective on DFF is demanded. More aggressive voltage scaling for DFF is expected to enable computing in ultra-low voltage circuits, such as [7], [39].

Chapter 3 SRAM Device and Circuits

Optimization toward Energy Efficiency in Multi- V_{th} CMOS

3.1 Background

Recently emerging micro-watt applications such as micro-sensor networks, handset electronics and implantable biomedical devices, etc., have placed their primary criterion on minimum energy consumption or high energy efficiency to prolong battery life time. To improve the energy efficiency, operating voltage (V_{DD}) in these applications is positioned near or below the threshold voltage (V_{th}), known as the near- or sub-threshold region. However, design of ultra-low voltage digital and memory circuits is highly required for achieving this ultra-low energy goal. Particularly, the design of ultra-low voltage SRAMs remains significantly challenging due to the additional constraints such as high sensitivity to process-voltage-temperature (PVT) variations, smaller cell stability, smaller voltage margin, and prevailing leakage current.

SRAMs can dissipate significant power and consume high energy in numerous applications, such as DSP, MCU and etc. Consequently, energy efficiency is a topmost parameter for SRAMs embedded in micro-watt systems. While numerous research works have been conducted for minimizing SRAM energy consumption, research on the utilization of multi- V_{th} devices for minimum energy-driven SRAMs has rarely been explored. The main challenge in the design of ultra-low voltage SRAMs with multi- V_{th} is to reduce leakage without degrading performance. In decoupled SRAM cells, higher- V_{th} devices are preferred in write paths and data storage to reduce leakage current, and lower- V_{th} devices are used in read paths to achieve better performance. However, this can generate excessively slower write operation than read operation if V_{th} of the devices in the write paths is too high compared to that of the devices in the read ports. This chapter thereby examines the

approaches to improve the energy efficiency of SRAMs with multi- V_{th} devices. Optimal device combinations will be analyzed for maximizing energy efficiency. We will also present the effects of various SRAM design techniques on enhancing the energy efficiency with multi- V_{th} devices. The rest of the chapter is organized as follows. In Section 3.2, we will analyze the energy consumption of SRAMs. The optimal V_{th} for maximum energy efficiency will be discussed in Section 3.3. Section 3.4 explains design techniques that can enhance the energy efficiency. Finally, we will make a summary in Section 3.5.

3.2 Analysis of SRAM Energy

Fig. 3.1 Simplified SRAM array diagram for energy analysis.

Energy efficiency is a paramount design criterion in emerging ultra-low power applications. Supply voltage scaling has been the most widely accepted method for energy efficiency improvement. SRAMs, however, require additional considerations such as array structures, active-switching, and leakage energy. Although dual- V_{th} and multi- V_{th} schemes have been utilized for power reduction [40], [41], minimum energy-driven device selections have been rarely visited. In this section, we will analyze SRAM energy minimization considering the option of multi- V_{th} devices. The functionality of all SRAMs is guaranteed by simulation, even at the condition of the lowest supply voltage.

3.2.1 SRAM Energy Modeling

The occurrence of a minimum energy operating point is determined by the correlation of power and performance. Energy consumption of an SRAM can be separated into two components: switching energy, also known as dynamic energy,

and leakage energy known as static energy. Fig. 3.1 shows a simplified SRAM array with highlighted critical parameters relevant to the energy analysis. The conventional 8T decoupled SRAM cell [13] is employed due to its popularity in ultra-low voltage SRAM design. The effect of an optimal device selection on the energy of SRAM peripheral circuits is insignificant compared to that on SRAM arrays. Therefore, peripheral circuits such as decoding blocks, column multiplexers for read and write operations, sense amplifiers and write drivers are excluded in this energy analysis.

The total energy (E_{total}) of the SRAM array can be expressed by

$$E_{total} = E_{switching} + E_{leakage} \quad (3.1)$$

where $E_{switching}$ represents the dynamic energy consumed by switching activities. $E_{leakage}$ is the static energy consumption coming from the leakage current in the SRAM cells. $E_{switching}$ is the summation of the switching energies during read operation and write operation, which can be expressed as below.

$$E_{switching} = P_{Read} \left(C_{RWL} \times V_{DD}^2 + k \times P_{Low} \times C_{RBL} \times V_{DD}^2 \right) + P_{Write} \left(C_{WWL} \times V_{DD}^2 + \frac{k}{m} C_{WBL} \times V_{DD}^2 \right) \quad (3.2)$$

where P_{Read} is the probability of read operation, P_{Low} is the probability of reading data '0' during read operation, and P_{Write} is the probability of write operation. Note that no switching activity occurs in the read bitlines when the read data is '1'. As shown in Fig. 3.1, the read energy is associated with the wordline capacitance (C_{RWL}) and the bitline capacitance (C_{RBL}). Note that multiple read bitlines (k) will be discharged during read operation due to the shared read wordline. Read data also affects the switching energy since the read bitlines are only discharged with the read data of '0'. Similarly, the write energy is primarily determined by the write wordline capacitance (C_{WWL}) and the write bitline capacitance (C_{WBL}). The switching write bitline capacitance is determined by the number of columns (k) and the multiplexing ratio (m). One write bitline in a pair switches regardless of write data. Therefore, the

write energy is independent of the write data.

The static energy ($E_{leakage}$) of the SRAM array is given by

$$\begin{aligned}
 E_{leakage} &= V_{DD} \times I_{Leakage} \times T \\
 &= V_{DD} \times N \times \left[I_{SN} \times e^{\frac{|V_{GS}| - |V_{thn}|}{nV_T}} + I_{SP} \times e^{\frac{|V_{GS}| - |V_{thp}|}{nV_T}} \right] \\
 &\quad \times (1 - e^{\frac{-V_{DS}}{V_T}}) \times T
 \end{aligned} \tag{3.3}$$

where, $I_{Leakage}$ is the total leakage current, N is the number of SRAM cells, I_{SN} and I_{SP} are technology scaling parameters for the NMOS and PMOS devices, V_{GS} is the gate-to-source voltage, V_{DS} is the drain-to-source voltage, V_{thn} and V_{thp} are the device threshold voltage of the NMOS and PMOS transistors, n is related to the sub-threshold slope, V_T is the thermal voltage, and T is the time to finish a computation. For simplicity, we assume that the sub-threshold current only consists of the drain current in the sub-threshold region.

3.2.2 Effects of Supply Voltage Scaling and Threshold Voltage on Energy Efficiency

As shown in Equation 3.1 ~ Equation 3.3, the energy consumption is highly sensitive to VDD and device threshold voltage. In the point of view of designers, the simplest method of improving energy efficiency is to scale supply voltage. Lowering VDD improves energy efficiency when the dynamic energy is dominant over the static energy. However, the static energy becomes significant when the supply voltage becomes near or below the device threshold voltage level. In this region, even though the leakage current still decreases by lowering VDD, the exponentially degraded performance quickly increases the overall static energy. As a result, the combination of the dynamic energy and the static energy generates an operating point that minimizes the total energy consumption. This point is generally found in the region where VDD is below the device threshold voltage.

Higher- V_{th} devices have been utilized in the design of ultra-low power SRAMs due to the exponentially decreased leakage current. The ultra-low power is obtained at the cost of degraded performance. However, compared to the effect of the supply voltage scaling on the energy efficiency, the effect of the threshold voltage on the energy efficiency is not straightforward. Increasing the threshold voltage decreases the amount of leakage current exponentially. However, increased threshold voltage degrades performance exponentially too. Consequently, the impact of the threshold voltage alteration on the static energy is determined by the ratio of the reduced leakage current to the increased operating delay. If the gain in the leakage reduction is larger than the loss in the performance, the overall energy efficiency improves by replacing with higher- V_{th} transistors. Contrarily, if the impact of delay degradation exceeds the gain in leakage suppression, the energy efficiency improves when lower- V_{th} devices are adopted.

3.2.3 Effects of Multi- V_{th} Devices on SRAM Energy

Circuit design utilizing multi- V_{th} devices has been widely used in digital circuits. Critical paths are preferred to be designed using lower- V_{th} devices while higher- V_{th} devices are favored in non-critical paths. The higher- V_{th} devices in non-critical paths reduce the leakage current and the lower- V_{th} devices maintain the required performance. However, this cannot be easily employed in conventional 6T SRAMs since the SRAM performance is directly related to the amount of leakage current. Instead, multi- V_{th} devices have been usually adopted to achieve balanced design parameters such as cell stability, performance, and write margin. Unlike the conventional 6T SRAMs, decoupled SRAM cells with separated read and write ports can accomplish the energy efficiency improvement by employing higher- V_{th} devices in the data storage and low- V_{th} devices in the performance limiting read port.

Fig. 3.2 illustrates a sample 8T dual-port SRAM cell designed with higher- V_{th} devices in the cross-coupled latch and the write access transistors, and lower- V_{th} devices in the read port. This is straightforward when considering that read operation is slower than write operation. In this case, the energy model described in Section 3.2.1 has to be modified. The energy equation for the switching energy remains the same while the leakage energy equation is written by

$$\begin{aligned}
 E_{leakage} &= I_{Leakage} \times V_{DD} \times T \\
 &= \left[\begin{aligned} &I_{SN_HV} \times e^{\frac{|V_{GS}| - |V_{thn_HV}|}{nV_T}} + I_{SP_HV} \times e^{\frac{|V_{GS}| - |V_{thp_HV}|}{nV_T}} \\ &+ I_{SN_LV} \times e^{\frac{|V_{GS}| - |V_{thn_LV}|}{nV_T}} \end{aligned} \right] \\
 &\quad \times (1 - e^{\frac{-V_{DS}}{V_T}}) \times N \times V_{DD} \times T
 \end{aligned} \tag{3.4}$$

Fig. 3.2 Schematic of an 8T decoupled SRAM cell with multi- V_{th} devices.

where I_{SN-HV} , I_{SP-HV} , and I_{SP-LV} are technology scaling parameters for the higher- V_{th} NMOS, higher- V_{th} PMOS and lower- V_{th} NMOS, V_{thn-HV} , V_{thp-HV} and V_{thn-LV} are the device threshold voltage of the higher- V_{th} NMOS, higher- V_{th} PMOS and lower- V_{th} NMOS. Compared to the previous leakage energy equation, three different types of devices (two types in NMOS and one type in PMOS) determine the cell leakage current. In addition to the leakage current, the time to finish a computation has to be rewritten as

$$T = \max(t_{read}, t_{write})$$

where t_{read} is the time to finish a read operation and t_{write} is the time to complete a write operation. If the write operation with higher- V_{th} devices takes longer time than the read operation with lower- V_{th} devices, t_{write} has to be used in the energy estimation. This indicates that increasing the threshold voltage of the higher- V_{th} devices over an optimal point quickly lose the energy efficiency improvement. In the following section, we will discuss the optimal SRAM cell design toward energy minimization using multi- V_{th} devices.

3.3 Minimum Energy-Driven SRAM Design Utilizing Multi- V_{th} Devices

Table 3.1 Parameter summary on energy analysis simulation

Items	Remarks
Technology	Commercial 65 nm CMOS
Array structure	256 rows $\times$ 128 columns
SRAM cell	8T decoupled SRAM cell
Devices	LVT, SVT, HVT
V_{th}	LVT: 0.28 V/-0.2 V SVT: 0.37 V/-0.31 V HVT: 0.61 V/-0.59 V
Read delay	From clock to RBL at $0.5 \times VDD$
Write delay	From clock to data flipping point
SRAM operation	Read probability = 0.5, write probability = 0.5

SRAM energy consumption is determined not only by supply voltage selection but also by device selection. It has been demonstrated that the minimum energy of an SRAM is found in the sub-threshold region. In this section, we will investigate the impact of device selection on minimum energy consumption. Table 3.1 summarizes the relevant design parameters used in the analysis. An SRAM array in commercial 65 nm CMOS technology is simulated over various combinations in device selection. We use three types of devices which are low- V_{th} device (LVT), standard- V_{th} device (SVT) and high- V_{th} device (HVT) available in the selected CMOS technology. Read delay is measured at points of crossing ' $0.5 \times VDD$ '. Write delay is measured as the delay from enabling to data flip point in the analysis. Instant read-after-write operation is not considered and it will be analyzed in Chapter 4. Process and temperature variations affect device characteristics. However, they are not included and this work will primarily focus on the effect of multi- V_{th} devices on SRAM energy minimization.

3.3.1 Analysis of SRAM Energy without Multi- V_{th} Devices

Fig. 3.3 Normalized energy of three SRAMs designed by three different device types (i.e. HVT, SVT and LVT). All transistors in one SRAM have the same V_{th} .

To minimize the leakage power consumption, SRAM cells have employed higher- V_{th} devices at the cost of performance degradation. However, the degraded performance caused by the selection of higher- V_{th} devices also affects energy consumption. Therefore, careful device selection has to be considered for improving energy efficiency using multi- V_{th} devices. Fig. 3.3 demonstrates the SRAM energy consumption designed by different device types sweeping supply voltage. When the supply voltage is in the super-threshold region, dynamic energy is dominant compare to leakage energy. Therefore, lowering supply voltage reduces overall energy consumption. As expected, the minimum point of each device selection is formed at a point where the supply voltage is around the threshold voltage of the devices. However, the minimum energy level using HVT (0.16) or SVT (0.12) is higher than that of using LVT (0.08), which explains that selecting SVT or HVT for improving power dissipation is not the best choice in terms of energy efficiency. This result can be explained as follows. Compared to LVT, SVT and HVT decrease leakage current and increase read delay. However, since the increase in the read delay is more significant than the decrease in the leakage, the SRAM arrays using SVT and HVT consume more energy overall.

3.3.2 Analysis of SRAM Energy with Multi- V_{th} Devices

Fig. 3.4 Impact of device selection on normalized energy of three SRAMs. Note that HVT devices are employed for read port in all SRAMs. Rest transistors in each SRAM cell adopt one device type.

Fig. 3.5 Normalized delay values of SRAM read and write operations designed with HVT devices.

Transistors with different threshold voltages are offered in recent CMOS technologies. This provides circuit designers with more opportunities to optimize circuits in performance, power, and energy. While higher energy efficiency can be achieved through a proper device selection, an undesirable device selection will

Fig. 3.6 Comparison of read delay (LVT) with write delay implemented with multi- V_{th} devices (SVT and HVT).

produce lower energy efficiency. Fig. 3.4 shows the impact of undesirable device selections on SRAM energy. HVT devices are employed in the read port to limit the overall performance. In this case, using SVT and LVT devices does not improve the energy efficiency because SVT and LVT devices in write paths dissipate more power without improving the overall performance.

In general, higher- V_{th} devices are employed in non-critical paths for reducing power while lower- V_{th} devices are adopted in critical paths for achieving high performance. Conventionally, read paths are considered as critical paths, limiting overall performance as shown in Fig. 3.5. Write paths are non-critical due to the faster operation speed than read paths. Therefore, lower- V_{th} devices have to be incorporated in read operation, and higher- V_{th} devices can be employed in write paths. However, as supply voltage decreases, the write speed with higher- V_{th} devices degrades faster than the read speed with lower- V_{th} , eventually making the write paths critical. In this case, overall energy consumption needs to be estimated carefully since the degraded critical path delay from write operation becomes more substantial. Fig. 3.6 explains the impacts of device selection on the critical path delay. Using the result of Fig. 3.3, LVT devices are used in the read port for enhancing the SRAM performance. When SVT devices are used in the write paths, the delay of the write paths is still smaller than that of the read paths using LVT

Fig. 3.7 Normalized energy of SRAMs utilizing three different device types (i.e. HVT, SVT and LVT) for data storage and write paths. Note that LVT devices are used in read port.

Fig. 3.8 Comparison of leakage current over various device combinations.

devices. As a result, using SVT in the write paths will decrease leakage current while maintaining the same performance, consequently reducing energy consumption. However, when HVT devices are adopted in the write paths, the delay of the write paths will be larger than that of the read paths at lower supply voltages. This occurs because the write delay increases exponentially from a higher supply level while the read delay starts to augment exponentially at lower VDD. Specifically, read delay is larger than write delay in a single V_{th} SRAM cell. For this

Fig. 3.9 Summary of normalized minimum energy consumption over various device combinations.

Fig. 3.10 Summary of normalized leakage current over various device combinations.

MTCMOS cell, read delay is larger initially. However, HVT devices have higher V_{th} than LVT devices. Thus current from the write paths degrades sharply when VDD is near the V_{th} of HVT devices (~ 0.6 V) whereas it is still super-threshold for LVT devices. The significantly degraded write performance will lose the benefit of utilizing higher- V_{th} for enhancing energy efficiency. Fig. 3.7 demonstrates simulated SRAM energy of various device combinations. As expected from Fig. 3.6,

the minimum energy of an SRAM array using SVT in write paths shows better efficiency due to the reduced leakage current. However, in case of using HVT in write paths, the minimum energy point is formed at the supply voltage of 0.4 V and the energy increases dramatically. Fig. 3.8 demonstrates the leakage current of the SRAM arrays under different device combinations. Although the selection of SVT in the write paths and LVT in the read paths has the lowest minimum energy level, it consumes the second largest leakage current.

Fig. 3.9 summarizes the normalized energy consumption of various device combinations. The energy variation of up to $6.24\times$ exists, which emphasizes the importance of careful device selection. Apart from energy efficiency, leakage reduction by device selection is equally important. SRAMs with less leakage power dissipation are more demanded in battery-powered applications, especially in sleep mode. The corresponding leakage currents of the devices combinations described in Fig. 3.9 are shown in Fig. 3.10. Note that the device combination for the highest energy efficiency (SVT(W)-LVT(R)) is not the best in terms of leakage. In addition, the device combination with the highest leakage (LVT(W)-LVT(R)) has the second highest energy efficiency. Therefore, careful device selections have to be made depending upon the system requirements. If an SRAM stays in an idle or sleep mode for majority of the life time, the leakage current becomes more significant than the energy efficiency during computational operations. However, the energy efficiency will be more significant if the SRAM workload becomes substantial. A design option is implementing the write paths with different device types. By separating the write access transistors and the latch with individual V_{th} devices, write delay and leakage current can be both improved. Similarly, MTCMOS methodology is also applicable to 6T SRAM cell. But the transistor size has to be carefully selected to maintain cell stability.

3.4 Design Techniques for SRAM Energy Efficiency Improvement Utilizing Multi- V_{th} Devices

Fig. 3.11 8T decoupled SRAM cells with leakage reduction techniques: (a) column-interleaved and (b) read buffer foot control

Fig. 3.12 Effect of column-interleaved scheme on SRAM energy. The reference design is using SVT devices in the write paths and LVT devices in the read path, which is also shown in Fig. 3.7.

As discussed earlier, SRAM energy is determined by multiple parameters such as

leakage current, dynamic current and critical path delay. Various SRAM design techniques have been proposed to improve the above parameters. In this section, we will explore the effects of various SRAM design techniques on energy efficiency under multi- V_{th} devices. Design techniques such as a column-interleaved scheme and a read buffer foot control scheme for leakage reduction, and boosting schemes for performance improvement will be considered. The energy overhead of utilizing the two techniques is negligible compared to the improvement they make. Other write performance boosting techniques such as data retention voltage collapsing and negative bitline scheme are also effective and applicable.

3.4.1 Effect of Power Reduction Techniques on SRAM Energy

Fig. 3.11 illustrates 8T decoupled SRAMs employing the column-interleaved scheme [42] and the read buffer foot control scheme [43]. In Fig. 3.11(a), the Column-Selected Line (CSL) is shared by SRAM cells in each column. During non-read operation CSL is held to VDD to eliminate the read bitline leakage from pre-charged RBL to CSL. During read operation, CSL in selected columns is pulled down to GND and RBL is conditionally discharged based upon the stored cell data. However, in unselected columns, CSL remains at VDD, which eliminates not only the bitline leakage in the read port but also the unwanted RBL discharging. Read buffer foot technique was proposed to reduce bitline leakage and enhance read bitline sensing margin. As Fig. 3.11(b) depicts, FOOT is shared by SRAM cells in each row. It can be either pulled-up to VDD to eliminate leakage current flowing to the read bitline or statically connected to GND to form a discharging path from read bitline to ground. During non-read operation, FOOT is connected to VDD to eliminate the leakage through the read port. During read operation, only FOOT in the selected row is pulled down to GND, and all RBLs are conditionally discharged based upon the data in the selected row. Compared to the column-interleaved scheme, the key advantage of the read buffer foot control scheme is to provide enhanced RBL sensing margin at low supply voltage. However, the column-interleaved scheme demonstrates better performance in point of power reduction since it eliminates the unwanted dynamic discharging as well as the RBL

leakage. Therefore, in this analysis, we will estimate the effect of the column-interleaved scheme on the overall SRAM energy. The combination of SVT in write paths and LVT in read paths as shown in Fig. 3.7 is also assumed in the analysis.

While the RBL leakage is avoided in both of the column-interleaved scheme (Fig. 3.11(a)) and the read buffer foot control scheme (Fig. 3.11(b)), the dynamic energy reduction is more significant in the column-interleaved scheme, which is primarily determined by the multiplex ratio. Although the selected column dissipates more power due to the discharging of CSL and the internal node in the read port, the elimination of discharging RBL in unselected columns improves the overall SRAM energy. Fig. 3.12 demonstrates the effectiveness of the column-interleaved scheme on energy efficiency. Simulation shows the energy reduction is proportional to the multiplex ratio. A multiplex ratio of 32 improves the energy efficiency by $\sim 5\times$ compared to the reference design whose device combination has the highest energy efficiency (Fig. 3.7). In addition, raising the multiplex ratio moves the minimum energy points to higher supply voltages, which is more desirable when considering the larger device variations at lower supply voltages.

3.4.2 Effect of Performance Boosting Techniques on SRAM Energy

Fig. 3.13 Simplified 8T SRAM schematic adopting boosted wordline scheme.

Fig. 3.14 Improvement of energy efficiency by boosting write performance.

Additional energy overhead induced by the boosting voltage generation is not considered in this simulation.

At a given SRAM array architecture, device selection has to be made for maximizing performance and minimizing leakage to achieve better energy efficiency. In Fig. 3.7, the highest energy efficiency is achieved by the combination of SVT in the write paths and LVT in the read paths. Although HVT in the write

paths can reduce the leakage more substantially, the exponentially degraded performance in write operation deteriorates the overall energy efficiency much faster. Boosted voltage schemes can be employed for enhancing write performance over read performance (Fig. 3.13). In this scheme, the voltage of WWL is boosted to a higher voltage than VDD. Consequently, the V_{GS} of the write access transistors increases and the write speed is enhanced accordingly. Fig. 3.14 demonstrates the change in the SRAM array after utilizing a boosted voltage scheme. As expected, the significant boosting in write performance eliminates the increase in the SRAM energy below the previous minimum energy point and improves the energy efficiency continuously even at lower supply voltages. The gain in the energy reduction expands as the supply voltage decreases. For example, 9.4× improvement was achieved at the supply voltage of 0.2 V. Fig. 3.15 summarizes the effectiveness of the boosted voltage scheme on various device combinations. The boosting voltage scheme is only useful in HVT(W)-LVT(R) and HVT(W)-SVT(R) whose write operation is slower than read operation at lower supply voltages. It is worth noting that the largest energy reduction is realized in HVT(W)-LVT(R) because the leakage in the write paths is the smallest and the performance is the highest. Compared to SVT(W)-LVT(R) whose energy efficiency is the highest before

Fig. 3.15 Comparison of normalized minimum energy consumption with write performance techniques.

performance boosting, HVT(W)-LVT(R) consumes 11% less energy. The relatively small improvement is due to the fact that although significant amount of leakage is reduced from the array by using HVT, the RBL leakage caused by the LVT devices dominates the overall leakage current. This limits the overall improvement in the energy efficiency.

3.4.3 Combination Effect of Power Reduction and Performance Boosting Techniques

Fig. 3.16 Improvement of minimum energy after adopting the column-interleaved scheme (Fig. 3.11(a)) and the boosted voltage scheme (Fig. 3.13). Multiplex ratio of 32 is assumed.

Fig. 3.17 Comparison of normalized SRAM minimum energy consumption.

To maximize the energy efficiency of an SRAM, both leakage reduction and

performance improvement need to be achieved at the same time. In this work, the maximum energy efficiency can be obtained from HVT(W)-LVT(R) after adopting the column-interleaved scheme (Fig. 3.11(a)) for power reduction and the boosted voltage scheme (Fig. 3.14) for performance improvement. Fig. 3.16 illustrate the SRAM energy of HVT(W)-LVT(R) after employing the aforementioned design techniques. Note that two design techniques improve the normalized minimum energy from 0.131 to 0.006 ($\sim 22\times$). Finally, the benefit of multi- V_{th} devices incorporated with the power reduction and performance boosting techniques are summarized in Fig. 3.17. When additional circuit techniques are not employed, the SRAM with the optimal device selection (SVT(W)-LVT(R)) consumes 31% of the SRAM energy designed solely by HVT devices. However, the optimal device selection moves from SVT(W)-LVT(R) to HVT(W)-LVT(R) after adopting the power reduction and performance boosting techniques. Consequently, the energy efficiency improvement of $33\times$ is achieved, which is larger than the energy saving from voltage scaling as shown in Fig. 3.7.

3.5 Summary

This chapter presents a comprehensive energy analysis of the SRAMs under multi- V_{th} devices. Although higher- V_{th} devices are preferred in the write paths for reducing power and energy consumption, a careful device type selection has to be considered to maximize the benefit of utilizing multi- V_{th} devices. Using higher- V_{th} devices in the SRAM write paths improve energy efficiency when VDD is in the strong inversion region where the write speed with higher- V_{th} devices is still higher than the read speed with lower- V_{th} devices. However, lowering supply voltage degrades the write speed faster than the read speed, eventually leading to slower write operation and losing the benefit of using higher- V_{th} devices in the write paths for power and energy reduction. Therefore, there exists a limitation in the V_{th} difference of the devices used in the write paths and the read paths. In this analysis, using the devices (HVT, SVT, and LVT) available in a commercial CMOS technology, the best device combination for energy minimization is to use SVT devices in the write paths and LVT devices in the read ports. We also explored the effects of several power reduction and performance boosting techniques on SRAM energy efficiency. After employing these techniques, the optimal device combination moves to HVT devices in the write paths and LVT devices in the read paths. This optimal combination improves energy efficiency by 33× compared to the device combination of HVT devices in the write and read paths.

Chapter 4 Design of an Ultra-low Voltage 9T SRAM with Equalized Bitline Leakage and CAM-assisted Energy Efficiency

4.1 Background

State-of-the-art DSP cores and advanced healthcare SoCs [44],[45] benefit from availability of on-chip SRAMs with substantially reduced power dissipation and improved energy efficiency. Integrated SRAMs play a crucial role in providing the required density, performance, power, and energy consumption of applications. By aggressively scaling supply voltage near or below transistor's threshold voltage, power and energy efficiency of SRAMs can be greatly ameliorated at the expense of performance. However, the vulnerability of SRAMs to PVT fluctuations makes reliable near- and sub-threshold operation extremely challenging in deep sub-micron CMOS technologies. Simultaneously, other design metrics such as stability, read/write margin, and leakage need to be carefully revisited for the reliable operation.

SRAMs have achieved ultra-low power/energy through supply scaling [16],[46],[47]. However, they suffer from various design issues mainly caused by reduced I_{on} -to- I_{off} ratio combined with large variations. Under severely scaled supply voltage, cell stability and bitline sensing margin of 6T SRAMs degrade dramatically due to the significant impact of disturbing current and bitline leakage. To handle it, an 8T differential SRAM cell [48] has been proposed to inject identical leakage current into the differential bitlines, eliminating the differential offset voltage from the leakage. However, in general, decoupled SRAM cells [16],[47] are preferable in weak-inversion regime to make the read SNM identical to the hold SNM. Moreover, the dedicated read port enables a faster read operation with no disturbing current to cell nodes.

Energy efficiency is a vital design metric for ultra-low voltage SRAMs. Although

voltage scaling decreases the switching energy quadratically, it deteriorates the operating frequency by several orders of magnitude. Accordingly, leakage energy accumulated in slow clock cycles would dominate the total energy in the deep sub-threshold region, leading to an energy contour shooting up [46]. To reduce the static energy, leakage current minimization techniques are desirable. In general logic circuits, adoption of HVT devices in non-critical paths is favorable to suppress the leakage. Another effective method to improve energy efficiency is suppressing leakage energy by eliminating idle gates or modules in the system, which is adopted by [49]. Leakage suppression is also attainable from algorithm level [50]. Among all the strategies for energy saving, leakage energy reduction is the first concern to improve energy efficiency.

In this work, we present several design techniques to foster an energy efficient SRAM in a wide range of supply voltages with the following features: 1) a decoupled 9T SRAM cell with an improved SNM compared to the 6T cell; 2) a 3T read port for equalizing RBL leakage and augmenting bitline swing; 3) utilizing MTCMOS technology for minimizing leakage in 6T write port and maximizing SRAM performance in read port; 4) a CAM-assisted circuit technique for improving the energy efficiency by boosting the write speed. The proposed circuit techniques are demonstrated by a 16 kb SRAM test macro (including the CAM) fabricated in a 65 nm CMOS technology.

4.2 Proposed SRAM Design Techniques for Ultra-low Voltage Operation

4.2.1 A Novel 9T SRAM Cell

Fig. 4.1(a) Proposed 9T SRAM cell implemented with HVT devices in write paths and LVT devices in read port. (b) Layout of the 9T SRAM cell based on 65 nm logic rules.

Fig. 4.1 depicts the proposed 9T SRAM cell and its layout. The cell consists of a 6T SRAM part (the write-access transistors with a cross-coupled latch) and a dedicated read port. The read port comprises three NMOS transistors ($M7$, $M8$ and $M9$) for realizing equalized bitline leakage and improving bitline sensing margin in a single-ended read bitline (RBL). The write access paths and the data storage latch are implemented with HVT devices for leakage reduction while the read port employs LVT devices for performance. The layout of the 9T cell occupies an area of $2.63 \times$

0.72 μm^2 based on logic design rules. A write operation is enabled by activating a write wordline (WWL) and completed when the data loaded at WBL and WBLB is written into Q and QB. A read operation starts by enabling a read wordline (RWL) and is followed by conditional RBL discharging. If Q holds logic '0', *M7* is turned on and discharges RBL to GND. If, on the contrary, Q stores logic '1', *M8* is activated and provides pull-up current from RWL (= VDD) to RBL, slowing down the discharging speed of RBL.

4.2.2 Analysis of Static Noise Margin and Write Margin

Fig. 4.2(a) 9T read SNMs compared with 6T and 10T SNMs with different voltages. (b) Distribution of 9T read SNMs at 0.4 V.

Decoupled SRAM cells, such as the 8T SRAM cell in [13] and the 10T SRAM cell in [17], have been widely accepted for SNM improvement. Eliminating the interference from read bitlines into cell nodes, such as the 8T cell and the 10T cell, makes the read-mode SNM equivalent to the hold-mode SNM. The read-mode SNM of the proposed 9T multi- V_{th} SRAM cell is compared to those of the conventional 6T cell and the 10T cell in Fig. 4.2(a). To investigate the impact of different V_{th} on SNM, the 6T SRAM cell is implemented with two device types. One is implemented with SVT devices and the other is implemented with HVT devices. Both pull-down NMOS transistors are over-sized by $1.67\times$. SVT devices with the same geometry as the 9T SRAM cell are utilized for the 10T cell with the assumption that no multi-threshold voltage option is adopted in [17]. The SNM values over the operating supply range are illustrated in Fig. 4.2(a). For the SVT cells, SNMs increase significantly with VDD and then slightly slows down in the super-threshold regime. The SNMs of the 6T and the 9T HVT cells, whereas, exhibit a more linear slope with supply voltage, which are far from saturation with increased VDD. This is partially caused by higher channel implant by HVT layer in this multi-threshold technology. The 9T cell shows a SNM of 52 mV at 0.2 V, improving the margin by 85.7% compared to the 6T HVT cell. At the nominal

Fig. 4.3 Comparison of write margins of HVT and SVT devices.

supply voltage, the SNM of the 9T SRAM cell is $1.13\times$ larger than the 10T SRAM cell whereas the difference of the two SNMs decreases at lower supply voltages. SNM Monte Carlo simulations for 3σ mismatch on top of the TT corner are conducted and the results are illustrated in Fig. 4.2(b). The 10k-point Monte Carlo simulations at $V_{DD} = 0.4$ V reveal that the proposed SRAM cell generates a mean SNM of 145 mV with a standard deviation of 17 mV. It provides a higher mean value with comparable variation than the 10T SRAM cell composed of all standard V_{th} (SVT) transistors.

For an SRAM cell, write margin is interpreted as the voltage headroom at write wordline for a successful write operation. Generally, it is determined by the drive strength ratio of the write-access transistors to the pull-up transistors. Simulated write margin of the 9T SRAM cell is plotted in Fig. 4.3. By sweeping supply voltage from 0.2 V to 1.2 V, the write margin increases from 34 mV to 320 mV with $9.4\times$ improvement. Utilizing SVT devices in the write paths can generate larger write margin due to its stronger writeability compared to HVT devices. Although the HVT devices in the 9T cell are relatively weak, they are employed in the entire write paths since compact cell layout for high-density integration and lower leakage is more important. The write failure in the 9T SRAM cell can be compensated by a CAM-assisted write performance boosting technique whose details will be explained in Section 4.2.3.

4.2.3 Bitline Leakage Equalization with the Worst Case of Leakage

Fig. 4.4(a) Conventional bitline sensing in the 8T SRAM [13]. (b) Concept of proposed 9T bitline sensing improvement by bitline leakage equalization technique.

During read operation, the voltage level of RBL is a function of VDD, device threshold voltage and leakage current, etc. At a specific VDD and a bitline length, the RBL level is highly affected by the amount of leakage current. Maximum bitline leakage occurs when the data in the unselected cells is all logic '0'. Similarly, minimum leakage current appears if the data pattern in the un-accessed cells is all logic '1'. Conventionally, a successful bitline sensing requires RBL for data '0' to

discharge much faster than that for data '1'. However, when variation in bitline leakage becomes comparable to cell read current, reliable detection of data '0' and '1' is difficult due to the small margin in the current to be sensed. In [16], it is shown that, in the worst case, the bitline level of data '0' could be even higher than that of data '1' due to the significant data-dependent bitline leakage particularly at ultra-low voltage.

The conventional bitline sensing problem caused by leakage at ultra-low voltage is illustrated in Fig. 4.4(a). In read operation, RWL is enabled and the RBL voltage forms depending on the accessed data. The pull-down strength for sensing '0' should be far higher than that of sensing '1'. After that, the simple sense amplifier (SA) consisting of two stages of invertors senses the voltage of each RBL without trigger timing. As illustrated in the bottom of Fig. 4.4(a), this requires the total of the cell current and the minimum leakage current ($I_{\text{cell}} + I_{\text{leak_min}}$) to be far larger than the maximum leakage ($I_{\text{leak_max}}$) for successful sensing. As the amount of leakage is comparable to the cell current and the leakage current varies from column to column due to different data pattern, this condition could not be always met.

To address this problem, we propose a bitline leakage equalization technique for single-ended read bitlines. Fig. 4.4(b) depicts the concept of the proposed bitline equalization technique utilizing the proposed 9T cell. In unselected cells, leakage current I_1 flows to GND through the device which is controlled by node QB when the data stored is logic '0'. Likewise, when the data is logic '1', leakage current I_2 flows to RWL (= GND) through the device controlled by Q. Accordingly, one of two devices connected to Q and QB ($M7$ and $M8$ in Fig. 4.1) is always turned on and the read access device ($M9$ in Fig. 4.1) is off. Consequently, two leakage paths have the same strength regardless of the stored data and the constant bitline leakage I_{leak} is formed. In Fig. 4.5, the RBLs are indicated as RBL for '1' with maximum leakage and RBL for '0' with minimum leakage. The pull down current for sensing '0' ($I_{\text{cell0}} + I_{\text{leak}}$) is always larger than that for sensing '1' ($I_{\text{leak}} - I_{\text{cell1}}$). This ensures that the RBL level for data '0' is always lower than that for data '1' and irrespective of the magnitude of I_{leak} . Thus, positive sensing margin could always be provided.

Fig. 4.5 Improved RBL swing and sensing window of 9T bitline at 0.2 V and $f_{CLK} = 50$ kHz with the worst case of leakage.

Sample simulated RBL waveforms (Fig. 4.5) show a drastically improved RBL swing in the 9T SRAM at $V_{DD} = 0.2$ V whereas the conventional 8T column (HVT(W)-LVT(R)) generates a negligible RBL swing. The proposed scheme improves the RBL swing by $4.6 \times$ at 0.2 V, 27°C, and 256 cells per bitline. Simultaneously, it also provides a wider sensing timing window, which is denoted by a double-side arrow. Note that the sensing timing window is defined as the time difference between the RBL of '0' and that of '1' measured when they cross $V_{DD}/2$. Since the trip point of our sense amplifier is $V_{DD}/2$, we used it as a reference level. With a frequency of 50 kHz, a sensing timing window of 1.5 μ s is achieved by the leakage equalization technique whereas nearly no sensing timing window is obtained in the 8T bitline. The RBL behavior of the 10T SRAM [17] is also captured in Fig 4.5. Apparently, the RBLs couldn't fully discharge at this frequency and they are too close to differentiate for sensing.

Variations of cell current and leakage current cause RBL swing to change as well as sensing problems. Fig. 4.6 depicts the distribution of RBL swing of the proposed 9T SRAM with 3σ local variation at the minimum operating voltage. With a mean

Fig. 4.6 Histogram of RBL swings of 9T SRAM at 0.2 V with the worst case of leakage from 10k-point Monte Carlo runs.

Fig. 4.7 Improved RBL swing with different numbers of cells and temperature. Typical corner is used in the simulation.

value of 53 mV, the RBL swing distribution from 10k-point Monte Carlo runs exhibits a longer right tail. Fig. 4.7 presents the simulated swing-to- V_{DD} ratio of the proposed 9T SRAM and the 8T SRAM at different temperatures and maximum numbers of cells per RBL (RBL lengths). In order to compare different bitcell topologies in terms of RBL length, we assume nominal process parameter values. In reality, accounting for within-die parametric variations, the effective number of cells per RBL degrades. The proposed 9T SRAM bitline can attach more cells due to the larger RBL swing as verified in Fig. 4.7. In the 8T SRAM bitline, only 512

cells can be attached for a sensible RBL swing. Note that a sensible swing should be at least a positive value. In the proposed SRAM, up to 1024 cells can be attached to the 9T bitline at 0.3 V and 80°C. The 8T bitline with 1024 cells generates a negative bitline swing at 80°C.

4.3 Proposed Energy Efficient Improvement Technique

4.3.1 Limitation of MTCMOS on SRAM Energy Efficiency

For a given SRAM structure, the energy efficiency can be optimized by minimizing leakage and maximizing performance. To realize it, the 9T SRAM cell consists of HVT devices in the 6T part and LVT devices in the read port. However, as explained in [51], this is not the best option in terms of energy efficiency, which is primarily due to the write performance degradation. Assuming 50% duty cycle, SRAM energy (E_{total}) can be written by

$$\begin{aligned}
 E_{total} &= E_{switching} + E_{leakage} \\
 &= C_{switching} \times V_{DD}^2 + I_{leakage} \times V_{DD} \times T \\
 \text{where } T &= 2 \times \max(t_{read}, t_{write})
 \end{aligned} \tag{4.1}$$

In the case, T is determined by t_{read} , using HVT in the 6T part reduces $I_{leakage}$ and improves the energy. For the other cases, when T is determined by t_{write} , the reduction in $I_{leakage}$ and the increase in T have to be carefully revisited.

Fig. 4.8 illustrates a write operation with data flipping. The write operation is

Fig. 4.8 Definition of data flipping delay and data full development delay. Difference of the data flipping delay and the full development delay substantially increases with scaling VDD.

Fig. 4.9 Read failure due to data non-full development in SRAM cell nodes.

divided into two stages, data flipping and data full development. After the data flipping, additional delay is required for the internal nodes, e.g. QB, to be fully developed. In skew conditions (e.g. MTCMOS cell, skew process corners), QB could move to high voltage very slowly. In this work, the delay till node crossing is defined as the data flipping delay. The delay till data full development (i.e. 90% of VDD) is defined as the data full development delay. The latter is more proper to measure the real completion of a write operation. Fig. 4.8 plots the difference between the data full development delay and the data flipping delay. It is clearly demonstrated that the delay difference between data flipping and full development sharply expands at ultra-low voltage operation.

In an SRAM circuit, the active clock duration is decided by the larger value between the write delay and the read delay. As supply voltage decreases, the write delay with HVT devices degrades faster than the read delay with LVT devices, eventually exceeding the read delay. In this scenario, the overall performance is limited by the slower write operation. To improve it, the flipping delay instead of the full development delay, can be adopted as write delay when no read-after-write operation is assumed. Fig. 4.9 shows the issue of the read-after-write operation. After the data flipping at Q and QB, QB rises slowly. When RWL is enabled, Q and QB have not been fully developed yet and the read operation could fail. The write data could be accessed only after additional clock cycles for full development.

Fig. 4.10(a) Read and write delays against scaling VDD at TT corner. (b) Read and write delays against scaling VDD at FN-1SP corner.

Consequently, the excessively degraded full development delay nullifies the energy efficiency by prolonging T even if significant leakage reduction is achieved with HVT devices in the 6T part. Fig. 4.10 depicts the read and the write delay of the 9T SRAM array at TT and FN-1SP corners, respectively. When the supply voltage is lowered below 0.6 V, the full development delay and the flipping delay deteriorate faster than the read delay (Fig. 4.10(a)). The former delay is $6.12\times$ of the latter delay at FN-1SP corner and VDD = 0.46 V, as shown in Fig. 4.10(b). The read delay is larger than the flipping delay when VDD is between 0.46 V ~ 0.64 V. In this simulated voltage range, data flipping is definite to occur within the read delay. Therefore, energy improvement can be obtained if the read-after-write issue could be eliminated by utilizing a faster delay (e.g. read delay/data flipping delay) as T .

To address the above issue and enhance energy efficiency, we propose a Content-Addressable-Memory-assisted (CAM-assisted) circuit for boosting write performance as well as compensating write failure.

4.3.2 Proposed CAM-assisted Write Performance Boosting Technique

Fig. 4.11 Data paths of write and read operations in the CAM-assisted SRAM circuit.

The slow-write-fast-read problem can be addressed in the architecture level [8]. A completion signal is asserted to alert the CPU when the write operation is finished, otherwise the CPU stalls for 2~3 cycle during write. Traditional bypass circuit implemented in SRAM utilizing registers to cache input data can also boost performance in the read-after-write case. However, firstly, the register cell can easily cost more than 16 transistors if a mainstream DFF style is adopted for ultra-low voltage operation. Secondly, large number of dependent MUXs and comparators are needed and could not be multiplexed. Therefore, it is beneficial to make the storage circuit, MUXs and comparators implemented with fewer transistors and in an area-efficient array-based way. In this section, we explain a circuit technique that can enhance write performance with this advantage.

Fig. 4.11 illustrates the operation of the proposed CAM-assisted technique. The SRAM comprises two main paths, an SRAM path and a CAM path. The SRAM

Fig. 4.12 Delay of four different operations of SRAM and CAM circuits.

path consists of a 16 kb 9T SRAM array (main SRAM array), decoders and data IOs. The CAM path is composed of a tiny 48b CAM array for storing addresses, a ring counter as an address pointer, an encoder, and a miniature SRAM array for storing write data. The CAM array (Addr.) and the SRAM array (Data) are implemented with LVT devices for faster read, write, and parallel search to conceal the slow full data development in the main SRAM array. The primary role of the CAM is to store most recent write addresses and data for possible subsequent read access till the data written into the main SRAM array is fully developed.

During write operation (Fig. 4.11 left), data is written into the main SRAM array (through the SRAM write path) and the miniature SRAM array (through the CAM write path). The write address is stored in the CAM array. The write address and the data in the CAM can be accessed in the succeeding cycles since the proposed CAM is implemented with LVT devices. During read operation (Fig. 4.11 right), the main SRAM array is accessed for normal read operation, and the CAM array is simultaneously searched using the read address as search data. If the read address is not found in the CAM array, the cells that are written in the preceding cycles couldn't be accessed. Thus, the selection signal from the encoder (Match = 0) will select the read data from the main SRAM array as the final data through MUX. If an address match occurs by a subsequent read-after-write operation, the encoder enables a wordline signal corresponding to the matched address. The wordline activates reading data from the SRAM array and later the data is sent to MUX.

Fig. 4.13 Circuit diagram of CAM array, search logics and miniature SRAM array.

Finally, using the selection signal from the encoder (Match = 1), MUX will select the data from the proposed CAM as the final data. In this case, the read data from the main SRAM array cannot be used as the final data since the data written in the previous cycle has not been fully developed due to the slow development speed of the latches using HVT devices. Therefore, the read data from the CAM should be selected as the final read data. Through this, the write performance is determined by the read operation or the data flipping delay, not by the slower full development delay. As a result, instant read-after-write operation for the same address is executable without slowing down the clock frequency for providing full data development in the main SRAM array.

Fig. 4.12 compares the delays of four different operations (i.e. SRAM read, SRAM write, CAM read and CAM write) to demonstrate the performance advantage of the proposed scheme. The delay of SRAM write is calculated by the full development time. As shown in Fig. 4.12, the delay of SRAM write is the largest whereas that of CAM write is the smallest. Since the CAM-assisted technique hides the slow SRAM write, the overall performance is improved from SRAM write to SRAM

Fig. 4.14 Timing diagram of SRAM array and CAM circuit during succession of write and read operations.

read. The performance improvement of 47.5% is achieved from simulation.

The schematic and the searching operation of the 10T CAM cell employed in this work adopts from [52]. The CAM cell comprises a 6T SRAM part and search logic circuits. Before search operation, the match line (ML) is precharged to VDD. A search operation starts by loading search data into the search lines. If the search data is different from the stored data, one of the search logic circuits would discharge ML to GND. Contrarily, if the search data is identical to the stored data, ML remains at a high voltage. The circuit diagram of the CAM-assisted circuit is described in Fig. 4.13. Conventionally, input of a CAM is data and output is a hit address. In this work, input is a read address and output is data. The CAM array is comprised of 4 rows (i.e. storing 4 most recent write addresses) and 12 columns (i.e. 12-bit address). The number of rows is mainly determined by the ratio of the data full development delay and the flipping delay. A ring counter is utilized to act as a

Fig. 4.15 Faster write completion in CAM array than SRAM array at different corners.

pointer for the CAM array. When a write operation is asserted, the pointer enables one row, writing the input address into the CAM array and the data into the miniature SRAM array. When a SRAM read operation is enabled, the address is loaded into the search lines (SL_{*i*} and SLB_{*i*}) of the CAM array. If the address is found from the CAM array, the corresponding ML(s) will be enabled. Otherwise, no ML is enabled and the search operation finishes. If multiple MLs are enabled, the encoder activates only one read wordline (CAM_RWL_{*i*}) corresponding to the most recent write operation. The activated wordline enables reading data through read bitlines (CAM_RBL_{*i*}) and sending the read data to MUX (Fig. 4.11). The number of rows in the CAM array can be estimated by the following equation if 50% clock duty cycle is assumed

$$N = \left\lceil \frac{\text{Data Full Development Delay}}{2 \times \text{Data Flipping Delay}} \right\rceil - 1 \quad (4.2)$$

If M in Fig. 4.14 is greater than 2, read operation is likely to fail in the subsequent read operation (50% duty cycle), which is addressed by the proposed CAM-assisted technique. To cover a case at FNTP corner (Fig. 4.10(b)), N should be at least $\lceil 6.12/2 \rceil - 1$, which is 3. In this work, we implemented 4 rows to provide a redundancy in N for real application.

The timing diagram of the proposed CAM-assisted SRAM is illustrated in Fig. 4.14. The data in the tiny SRAM (CAM_Q/QB) develops much faster than that in the main SRAM array (SRAM_Q/QB). In the subsequent read operation, input address in the search lines (SLs) keeps the corresponding ML high, and accordingly quickly generates CAM_RWL and CAM_RBL due to the LVT devices and the small load. The other read path through RWL generates SRAM_RBL with a larger delay and, in the worst case, it generates a read failure. Fig. 4.15 manifests that the full development delay of the CAM is always smaller compared to the main SRAM array at all corners. Simultaneously, the full development delay of the CAM is also shorter than the read delay of the main SRAM array, making the read paths critical.

4.4 Test Chip Implementation and Measurement

Fig. 4.16 Measured (a) leakage current of the test chip and (b) write, read and average power at maximum operating frequency.

Fig. 4.17 Measured (a) read access time and (b) improved operating frequency of the CAM-assisted SRAM.

The main SRAM array is organized with $256\text{ words} \times 4\text{ bits} \times 4$, which occupies an area of $169\text{ }\mu\text{m} \times 195\text{ }\mu\text{m}$ (including power rails in rows and columns). It is divided into 4 sub-blocks and each sub-block is composed of 16 columns, sharing one IO. The CAM array is configured with 4 rows and each row has 12 CAM cells for storing addresses and 4 SRAM cells (LVT) for storing write data. The proposed CAM circuit occupies $1061\text{ }\mu\text{m}^2$ (not including interconnections), which is at least 60% smaller than the DFF-based design in our estimation. It causes an overhead

Fig. 4.18 Measured energy of SRAM only and the CAM-assisted SRAM.

approximately as 3% of the SRAM array area. The overhead will be less at a higher SRAM array density since the number of rows is mostly determined by a single cell. The energy dissipation by the proposed CAM circuit occupies a very small portion of the overall consumption. Simulation shows that the CAM energy per read with search operation is 59 fJ at 0.4 V with frequency of 1MHz. To be more flexible, data from CAM_RBL and SRAM_RBL can bypass the MUX for separate measurement.

A 16 kb SRAM test chip is fabricated in a commercial 65 nm CMOS technology with a nominal VDD of 1.2 V. Fig. 4.16(a) shows the experimental results of the leakage current. At 27°C, the leakage current of the test chip changes from 139.3 μ A (1.2 V) to 1.4 μ A (0.1 V). When temperature goes to 100°C, it becomes 305 μ A and 4.5 μ A, respectively. Power of read and write operation is measured at the maximum operating frequency (Fig. 4.16(b)). The read power is larger than the write power due to the precharging and discharging current in the read bitlines. The average power is measured in the supply range of interest, assuming equal probability of performing read and write operations. It changes from 146 μ W at 0.6 V to 4.12 μ W at 0.32 V. Fig. 4.17(a) verifies that the CAM circuit can provide a shorter read access time by 25.8% at 0.6 V and 7% at 0.38 V compared to the SRAM without the CAM. Below 0.38 V, read from CAM takes more time due to slow search operation at ultra-low voltage. The operating frequency of the CAM-assisted SRAM is depicted in Fig. 4.17(b). Around the critical voltage of 0.6 V, the

CAM circuit speeds up clock frequency of the main SRAM to 40 MHz. The maximum operating frequency at $V_{DD} = 0.4$ V is boosted to 5 MHz. The SRAM performance is therefore improved by 42.6% and 66.7% at 0.6 V and 0.4 V, respectively. The plot of energy per operation is shown in Fig. 4.18. The SRAM consumes an energy per operation of 3.47 pJ at $V_{DD} = 0.42$ V. Thanks to the CAM-assisted circuit, a minimum energy per operation (E_{min}) of 2.07 pJ is achieved and the energy efficiency is consequently improved by 40.3%. Averagely, the energy

(a)

(b)

Fig. 4.19(a) Readout waveforms capture at 0.26 V. (b) Die micro-photograph.

Table 4.1 Design metric comparison with various ultra-low voltage SRAMs.

	A-SSCC 2012 [53]	JSSC 2013 [54]	JSSC 2013 [55]	This Work
Technology	65 nm	65 nm	65 nm	65 nm
Density	128 kb	2 kb	32 kb	16 kb
Transistor count	8T	9T	7T	9T
Cell size	N.A.	1.24 x 2.31 μm^2	N.A.	2.63x0.72 μm^2
V_{DDmin}	0.37 V	0.28 V	0.26 V	0.26 V
Access time	N.A.	4.55 μs (0.3 V)	0.55 μs (0.26 V)	0.85 μs (0.26 V)
Leakage current	N.A.	0.05 μA (0.4 V)	N.A.	1.4 μA (0.1 V)
Min. energy (E_{min})	21.2 pJ	0.57 pJ	5.6 pJ	2.07 pJ
Normalized E_{min}	162 aJ/b	278 aJ/b	171 aJ/b	126 aJ/b

efficiency in the supply range of 0.38 V to 0.6 V is enhanced by 29.4%. The test cells are fully functional down to 0.26 V with the maximum operating frequency of 100 kHz (27°C). The read access time of the SRAM is measured as 0.85 μ s at 0.26 V. The CAM circuit achieves an average improvement of 18.7% in the read access time between 0.38 V ~ 0.6 V. It lowers the minimum read voltage further from 0.26 V to 0.23 V. The test chip micro-photograph with waveform capture is shown in Fig. 4.19. Table 4.1 compares the test chip with various ultra-low voltage SRAM circuits. Among the SRAMs, this work achieves the lowest minimum energy when it is normalized with respect to density.

4.5 Summary

Leakage and energy efficiency are primary concerns for ultra-low voltage SRAM design. This chapter presents several circuit techniques to implement an energy efficient SRAM with reliable read operation under ultra-low voltage. The proposed 9T SRAM cell with equalized bitline leakage fosters SRAM read operation at ultra-low voltage, achieving read access time of 79 ns and 0.85 μ s at 0.4 V and 0.26 V, respectively. To further reduce the static energy, MTCMOS technology is utilized to reduce the leakage in the SRAM array. While HVT devices in the 6T part reduce leakage, they degrade write performance significantly at low voltage. This nullifies the energy efficiency improvement in the near- or the sub-threshold region. To tackle this issue, we proposed a CAM-assisted write performance boosting circuit to speed up clock frequency. The test chip shows an average energy efficiency improvement of 29.4% with the aid of the proposed circuit technique. Consequently, the energy efficiency is improved by 40.3% with the minimum energy per operation of 2.07 pJ at 0.4 V. The measurement results prove that the proposed techniques are good circuit solutions for ultra-low voltage and energy efficient applications.

Chapter 5 Design of an Ultra-low Voltage Disturb-suppressed Dual-port SRAM

5.1 Background

Contemporary computing platforms with big data enable unprecedented interaction between human and computational resources. The ubiquitous computing necessitates high computing power with multi-core processing units and multi-port SRAMs. Dual-port SRAMs are accordingly highly demanded, even by energy-constraint applications, whose low energy consumption mainly attributes to ultra-low voltage operation.

Conventional 8T dual-port (DP) SRAM cells are derived from standard 6T single-port (SP) SRAM cells. Consequently they inherit the weakness of the 6T SRAM cells at low voltage operation like poor cell stability, reduced read-ability and write-ability, which impedes voltage scaling. The above issues are exacerbated because the 8T DP SRAM cell has two more access transistors leading to larger disturbance, especially at the common-row-access mode.

As Section 2.2.2 analyzes, the worst case of read-ability, write-ability and cell stability occur at the common-row-access mode where the two SRAM cells sharing the same wordlines are accessed via the two designated ports in one clock cycle. Under this circumstance, both wordlines for port access are enabled, which makes the total four access transistors exposure to noise. For a cell to be read, the storage nodes will suffer a disturb current from the other port which impedes the discharging of the read current and results in read-ability degradation or read failure. For a cell to be written, the process of the data flipping will be interfered with a disturb current from the other bitlines, which degrades the write-ability. For unselected cells along the wordlines, disturb current will be injected to the storage nodes from two directions so the cells are extremely susceptible to noise, which is known as cell stability downgrading.

These limit the minimum operating voltage ($V_{\min}$) which is generally higher than that of the 6T SP SRAM cell. Various circuits have been proposed to either improve read-/write-ability of 8T DP cell or reduce read-write disturb under the common-row-access circumstance. As Section 2.2.3 discusses, a priority row decoder with bitline shifter has been proposed to circumvent the common access mode for enhancing cell stability [30]. A wordline-voltage-adjustment system has been utilized to improve read and write against PVT variation [31]. It is obvious that the 8T DP SRAM cell has to employ assisting circuits to accommodate challenges from aggressive voltage and technology scaling.

In this chapter, we propose a 12T DP SRAM cell with 2 decoupled read ports for better read-ability, write-ability and cell stability without any assisting circuit. In addition, a virtual-ground scheme and hierarchical bitlines are deployed to suppress the leakage current from read bitlines and to improve the performance, which further improves $V_{\min}$.

5.2 Proposed 12T DP SRAM Cell

Fig. 5.1(a) Schematic of proposed 12T dual-port SRAM cell. (b) Layout of the 12T dual-port cell.

Dual-port SRAMs boost computation performance and throughput by doubling the number of simultaneous memory access. For the conventional 8T DP SRAM cell, port A and B are accessed by exclusive address and operation instruction. Each port consists of their corresponding wordline (WL) and a pair of bitlines (BL and /BL). In the common-row-access mode, the selected cell is inevitably disturbed through the second activated WL. Thus, the width of the 2 NMOS drive transistors has to be further expanded (e.g. $\times 2.7$) to maintain the cell stability. However, the upsized 8T DP cell still has limitations for ultra-low voltage operation, which will be analyzed in Section 5.3. Decoupled DP SRAM cells are promising solutions for that.

5.2.1 12T SRAM Cell Design

Fig. 5.1(a) portrays the proposed 12T DP SRAM cell. The proposed cell decouples read paths from write paths by implementing exclusive read port A ($M1 \sim M2$) and read port B ($M3 \sim M4$). The read wordlines (RWLA and RWLB) control the access to the two read ports by switching on the read access transistor $M1$ or $M3$. The RBLA and the RBLB are precharged to VDD in prior. During read cycle, they are left floating for data evaluation based upon the fighting between the read current in the selected cell and the leakage current from the unselected cells along the bitline. The conditional discharging of the read bitlines (RBLA and RBLB) is manipulated by VGND employed to suppress leakage. The corresponding VGND terminal is pulled down to ground for the selected column to foster a read operation whereas it is precharged to VDD to reduce leakage to the unselected columns. During read, the voltage level of RBLB represents the opposite value of node Q, hence it is connected to a global bitline via a PMOS transistor for data inversion. Write operation for the 12T SRAM cell is activated by write wordlines (WWLA and WWLB) and followed by data flipping in the storage nodes Q and QB, which is exactly the same as the 8T single-port SRAM cell. By separating the read paths from the write paths, read-write disturb is significantly relaxed and various design metrics such as stability and read-write margins are improved. This will be further discussed in Section 5.3. The proposed cell eliminates the necessity of over-sizing of the pull down devices while improves the key design metrics.

The layout of the proposed DP SRAM cell is illustrated in Fig. 5.1(b). Both of the WBLs and the RBLs run with the second metal layer in vertical direction. The RWLA and the RWLB maintain the third metal layer while the WWLA and WWLB go with the fifth metal layer. The power-line (VDD and VSS) and the virtual ground terminals also run with the second metal layer, resulting in eleven vertical tracks in total. The dimension of the 12T SRAM cell in the 65 nm technology is $2.75 \mu\text{m}^2$. To align with row decoder circuit, the height of the cell remains as $0.72 \mu\text{m}$. In our design, as the electrical β ratio is reduced from 2.7 to 1, the area overhead caused by the read ports can be partly compensated.

5.2.2 Implementation of Virtual Ground for Bitline Leakage Reduction

Fig. 5.2(a) Leakage problem in conventional 2T read port. (b) Read bitline leakage suppression by implementation of virtual ground technique.

Leakage current is detrimental in energy constrained applications as it consumes energy all the time, irrespective of data activity and event trigger. To make things worse, the aggregate leakage component at ultra-low voltage can offset the energy efficiency with voltage scaling and deflect the energy per operation from the optimum point. This problem becomes even more severe in the proposed 12T DP SRAM circuits when the leakage paths may double than the 8T SP SRAM.

During non-read cycles, the read bitlines are conventionally precharged to high voltage while the source terminals of $M2$ and $M4$ are normally grounded. However, this creates leakage current paths from the read bitlines to ground for unselected cells. Fig.5.2(a) illustrates the leakage current injection problem. Assume the very first cell is accessed through read port B. Although the other cells sharing the same

Fig. 5.3 Implementation of virtual ground technique.

read bitlines are switched off for read access, leakage paths form as long as voltage difference exists. Accordingly, every unselected cell suffers leakage current in two directions, one from RBLA to ground and the other from RBLB to ground. In addition, leakage current in standby mode is unwelcome in terms of power and energy consumption.

To minimize the leakage from the read ports, our design leverages a virtual ground technique (VGND) [43] by controlling the source voltage of $M2$ and $M4$ to suppress the leakage current. Unlike the row-wise implementation, this design adopts column-based virtual ground control to prevent bitline discharging in read for unselected columns. Fig. 5.2(b) illustrates the technique. Only during read operation, the corresponding VGNDs of the selected columns are grounded. Otherwise, it is pulled-up to VDD to eliminate the leakage injection. Fig. 5.3 shows the control circuit to implement the virtual ground scheme. When a port of the column is selected for read, COL_SEL and RD are enabled simultaneously. NOP is deactivated to indicate the circuit is in non-standby mode. The control pattern causes the level of SELECT to rise, which turns off the precharged PMOS device and switches on the transmission gate. Thus, VGND is synchronized with the inversion of clock (CLKB), which means VGND discharges to ground at the high phase of the clock. When a port of the column is unselected for read, the PMOS device provides current to pull up VGND.

5.3 Disturb Suppression of 12T DP SRAM in Common-Row-Access Mode

Fig. 5.4(a) Cell stability issue in common-row-different-column access. (b) Cell stability issue in common-row-common-column access.

Dual-port (DP) SRAMs have various access modes based upon row and column selection. Worst case disturb occurs when two selected SRAM cells are in the same row, which is called common-row-access mode. Fig. 5.4 illustrates the half-selected circumstance at the common-row-access for the conventional 8T DP SRAM. In Fig. 5.4(a), two DP cells are accessed from two designated ports, respectively. The cells sharing the same wordlines are all half selected by enabled wordlines. In Fig. 5.4(b), one DP cell is accessed by the two ports simultaneously. In both cases, the selected cells are disturbed by the current from bitlines to cell nodes through four activated access transistors. Simultaneously, the half-selected cells also suffer from disturb as the conventional SP SRAMs. Therefore, every 8T DP cell in the selected row is exposure to noise and subjected to cell stability issue. This chapter discusses the challenge of disturb in the common-row-access mode.

5.3.1 Analysis of Disturb Occurrence Probability

Regardless of 8T DP cell or 12T DP cell, the worst-case cell stability occurs when the data storage nodes are exposure to disturb current induced by both ports. Consequently, Fig. 5.5 summarizes SNMs in different situations with respect to

Fig. 5.5 Comparison of worst SNM scenarios in conventional DP SRAM cell and proposed DP SRAM cell.

operation in each port including read (R), write (W), half-selected by read (HR) and half-selected by write (HW). For the conventional 8T DP cell, SNM degrades most as long as two wordlines are enabled simultaneously except for write operation, where SNM is meant to be destroyed. Accordingly, 2Read, 1Read1Half-Selection, 2Half-Selection in com-row-access mode are all worst SNM cases. As indicated in Fig. 5.5(a), the probability of worst SNM for the 8T DP SRAM is 9/16.

The proposed 12T SRAM cell substantially decreases the probability of suffering stability degradation thanks to the decoupling of read and write ports. In the 12T cell, read operation does not affect SNM as it is isolated from the storage nodes. Therefore, only the situation where the cell is simultaneously half-selected by write can impose the worst-case disturb (Fig. 5.5(b)). This minimizes the disturb occurrence probability from 9/16 to 1/16, achieving an improvement of 88.9%.

Although the patterns of the worst case SNM are very complicated for analysis, most of the worst case patterns can be categorized into two classes, read disturb and write disturb. Read disturb is the situation where disturb occurs during read operation. Similarly, write disturb suggests disturb happens during write operation. Note that the other half-selection cases can be considered as dummy read operation, which will be treated as read disturb for analysis. Next two sections will provide insightful investigation on these two circumstances.

5.3.2 Analysis of Read Disturb

Fig. 5.6(a) Illustration of read disturb in the 8T DP SRAM cell. (b) Read disturb suppression in the 12T DP SRAM cell.

The read disturb of the conventional 8T DP SRAM cell is described in Fig. 5.6(a). Suppose port A are selected for read and port B are either selected or half-selected. The precharged bitlines for port A are left floating for data evaluation. As the access NMOS device is strong for passing '0', the read operation is dominated by the discharging capability of the precharged bitline through the storage node with '0'. As port B is simultaneously selected or half-selected with continuously precharged bitlines, it acts as a dummy read operation. The corresponding discharging current, as depicted in red in Fig. 5.6(a), injects to the same data storage node. Therefore, it

Fig. 5.7 Simulated waveforms of read disturb for the 8T DP cell and the 12T DP cell at $V_{DD} = 0.4$ V, FNSP corner. Note that the data in the 8T cell flips due to the read disturb whereas the data in the 12T cell maintains.

impedes the discharging of the bitline in Port A by raising the voltage of the storage node and increasing the load of the pull-down NMOS in the latch. This degrades read speed and cell stability, and can eventually result in read failure or data flipping [57].

The read disturb of the proposed 12T DP SRAM cell is depicted in Fig. 5.6(b). The read current, unlike the 8T DP cell, comes from the read bitline and discharges to ground through the read port with no interference with the data storage nodes. Although the read disturb current exists, the voltage of the node is not raised as much as that in the 8T DP SRAM cell because the destructive read operations are ameliorated from two ports to one port. Hereby, the cell suffers a much less possibility of data flipping.

Fig. 5.8 Comparison of read SNMs of the 8T DP SRAM and the 12T DP SRAM.

Fig. 5.7 presents simulated waveforms of the data storage nodes for both cells at FNISP corner using a 65 nm technology. Note that the NMOS drive devices are upsized by 2.7 $\times$ in the 8T cell. When $V_{DD} = 0.4$ V, the data in the 8T DP SRAM cell flips due to the read disturb from Port B whereas the 12T DP SRAM cell maintains the original data successfully with the presence of the read disturb. The SNMs in the common-row-access mode of the 12T DP SRAM and the 8T DP SRAM are compared in Fig. 5.8. An SNM of 58 mV is observed in the 12T DP SRAM cell at 0.4 V, which is improved by 26% compared to the 8T DP cell. The SNM of the proposed DP SRAM is greater than that of the conventional DP SRAM when the supply voltage is below 1 V. The data validates that the 12T DP SRAM cell has stronger immunity to read disturb compared to the 8T DP cell at near- or sub-threshold region, which is beneficial for ultra-low voltage operation.

5.3.3 Analysis of Write Disturb

(a)

(b)

Fig. 5.9(a) Write disturb illustration of the conventional 8T DP SRAM. (b) Write disturb suppression from the 12T DP SRAM.

The write disturb issue of the conventional 8T DP SRAM cell is described in Fig. 5.9(a). Suppose port A is selected for write and port B is either selected or half-selected. As NMOS transistors is strong for passing '0', the write operation is driven by the discharging capability of the storage node with '1' as Fig. 5.9 indicates. As Section 5.3.2 explains, a dummy read operation conducts in port B where the disturb current generates from the constantly precharged bitline. The current is injected to the storage node and impedes the voltage discharging. In

Fig. 5.10 Circuit of hierarchical write bitline.

addition, as the single bitline links to a great number of SRAM cells which can range from hundreds to tens of hundreds, the associated large capacitive load slows down the discharging speed at ultra-low voltage. Thus, the relative long-time write disturb can result in a write failure [57].

The proposed DP SRAM utilizes hierarchical write bitline scheme to reduce the bitline capacitance and ease the discharging. Although the write disturb pattern is very similar to that of the 8T DP cell, the bitline capacitance is reduced for faster disturb current discharging as depicted in Fig. 5.9(b). Fig. 5.10 presents the hierarchical bitline circuit. The 256 SRAM cells are linked to global write bitlines whereas each 64 SRAM cells are organized by local write bitlines. Each local bitline pair is allocated with individual precharge devices. The global bitlines access the local ones via transmission gates designated for each sub-block. Hereby, the associated bitline capacitance is mainly related to the 64 cells instead of 256 cells.

5.4 Measurement Results

Fig. 5.11 Architecture of the 65 nm test chip.

The proposed 12T dual-port SRAM is fabricated in 65nm CMOS technology. Fig. 5.11 presents the architecture of the SRAM test chip. The 16 kb SRAM array is configured by 256 words $\times$ 32 bit $\times$ 2. Each column is divided into 4 sub-blocks to implement the hierarchical write bitline. As Fig. 5.11 depicts, the dual-port SRAM has two access interfaces with dedicated peripheral circuits such as control logics, decoders, read-out circuits, I/Os, etc. The layout of the virtual ground (VGND)

Fig. 5.12(a) Measured leakage current. (b) Measured read access time.

(a)

(b)

Fig. 5.13(a) Measured power consumption of read and write. (b) Measured energy per operation.

(a)

(b)

Fig. 5.14(a) Micro-photograph of the test chip. (b) Captured waveforms of the RBL at VDD = 0.4 V.

circuit is implemented in vertical direction to align with the bitlines. With respect to the peripheral circuits, the layout is implemented as symmetric as possible. The test chip occupies an area of $398 \times 385 \mu\text{m}^2$ while the dual-port SRAM cell has a dimension of $3.82 \mu\text{m}$ by width and $0.72 \mu\text{m}$ by height.

The chip measurement has been conducted at the common-row-access mode and all data except leakage current is collected under this circumstance. The test chip is functional from 1.2 V to 0.4 V. Fig. 5.12(a) presents the measurement result of leakage. The leakage current decreases with supply voltage and hits a number of 7.6

μA at the minimum operating voltage while it is $63 \mu\text{A}$ at the nominal voltage. The simultaneous read operations are observed from the test chip. The read current is recorded with the maximum frequency and an equal probability of read '0' and read '1'. The total power is obtained by multiplying the current and the voltage. Energy per operation is the power consumed in an operation cycle. Fig. 5.12(b) describes the measured read access time (including I/O delay), which is the larger value of the delays from single port operation. The read access time varies from 6 ns at 1.2 V to 580 ns at 0.4 V. The plot exhibits a trend of exponential increasing with scaled VDD. Fig. 5.13(a) depicts the read and the write power consumptions with a maximum frequency clamp of 50 MHz due to the limitation of the equipment. The slope of the read power becomes steep below 0.7 V because the corresponding maximum frequencies are not affected by the clamp value. At 0.4 V, the test chip dissipates a power of $5.8 \mu\text{W}$ with a frequency of 350 kHz for successful read operations through the two ports in a common row. The write power shows a similar trend to the read power curve. The average current for write is calculated based on different write patterns, such as write '0'–write '1', both write '0' and both write '1'. The 16 kb SRAM consumes a write power of $4 \mu\text{W}$ with the same clock frequency. Fig. 5.13(b) presents the energy per operation of the proposed dual-port SRAM macro. The energy contour decreases with voltage and reaches a minimum point. It shoots up again with further voltage scaling because the performance degrades too much and the leakage component deteriorates the total energy efficiency. The minimum energy of 8.7 pJ is measured at 0.48 V. Fig. 5.14(a) shows the micro-photograph of the test chip and Fig. 5.14(b) captures the waveforms when the chip is working in the common-row-access mode with the minimum supply voltage.

5.5 Summary

This chapter presents a near-threshold dual-port SRAM circuit with suppressed read and write disturb at the common-row-access mode. To ameliorate disturb occurred in the conventional design, a novel decoupled SRAM cell is proposed to reduce the probability of disturb occurrences and eliminate the impact of the read disturb. A hierarchical write bitline scheme is implemented to boost write operation and combat with the write disturb. To minimize the bitline leakage, the proposed SRAM circuit utilizes a virtual ground technique in the read ports. The 12T dual-port SRAM is fabricated in a 65 nm CMOS technology and achieves a minimum operating voltage of 0.4 V at the common-row-access mode. This is the lowest voltage among all reported designs to the best knowledge of the author.

Chapter 6 Design of an Ultra-low Voltage, Energy-Delay Efficient Charge-Pumped DFF

6.1 Background

Fig. 2.13 Schematic of Transmission-Gate FF (TGFF).

Fig. 2.16 Schematic of Adaptive-Coupling FF (ACFF) [37].

D-Flip-Flop (DFF) as fundamental unit in integrated circuits can account for substantial random-logic power and area [37]. It is widely adopted in various applications such as polar code decoders, wireless channel equalizers, and seizure classification processors [58]-[60]. In emerging energy-constrained systems, minimum power/energy consumption has been insistently pursued. To attain it,

Fig. 2.17 Schematic of Static-Single-Phase Contention-Free FF (S²CFF).

supply voltage is usually positioned near or below the threshold voltage of transistor for minimum energy expenditure. However, aforementioned energy-efficient systems typically suffer drastic performance degradation and variation problems. Therefore, energy-delay-efficient DFF circuit is most desirable for near-/sub-threshold DFF applications.

To achieve reliable operation with energy-delay efficiency over a wide range of supply voltages, a DFF must satisfy the following requirements: 1) fully static, as static circuits are more tolerant to PVT variations especially at ultra-low voltage; 2) single-phase clocking, since the toggling of internal clock inverter incurs large power consumption; 3) less occurrence of setup time and hold time violation; 4) minimum device count compared to traditional DFFs for less area and less leakage.

Section 2.3.2 and Section 2.3.3 analyze 2 conventional and 2 emerging DFF circuits. To refresh, the pros and cons of the transmission-gate FF (Fig. 2.13), the adaptive-coupling FF (Fig. 2.16) and the S²CFF (Fig. 2.17) are rephrased here. The related schematics are reposted in this chapter for convenience. The mainstream transmission-gate FF (TGFF) is suitable for near-/sub-threshold operation due to its robustness at low voltage scenarios. However, the main challenge of the TGFF for energy-saving applications is the large power dissipation and low efficiency in energy. Specifically, the requirement of local clock buffer increases its power consumption and area overhead. To eliminate the clock buffer, a 22-transistor single-phase-clocking adaptive-coupling FF (ACFF) circuit has been proposed [37].

By deploying a differential structure with adaptive coupling scheme, it eases data transition, saves power and exhibits better energy efficiency than the TGFF. Despite its superior in energy efficiency, the ACFF typically works at super-threshold region (> 0.75 V), which is not fully qualified for ultra-low voltage operation [37]. Recently, a static single-phase-clocked 24-transistor S^2 CFF [38] has been proposed for low power applications. As Section 2.3.3 analyses, this DFF eliminates the clock buffer to improve energy efficiency and enhances the robustness of the circuit by utilizing keepers and glitch prevention technique. However, the transistor count is relatively large which can result in area and leakage inefficiency.

In this chapter, we present an ultra-low voltage and energy-delay efficient DFF for near-/sub-threshold applications. The transistor count of the proposed DFF is further decreased from 22 to 16 to save power and area, which is the minimum among all existing static style DFF circuits to the best knowledge of the authors.

6.2 A Novel Sub-threshold DFF

In order to achieve energy-delay efficient operation across a wide range of supply voltage, a DFF should have the following features: 1) static operation to against PVT variations; 2) single-phase clocking to suppress power consumption on the internal clock buffer; 3) minimum or less occurrences of setup time and hold time violations; 4) minimum or less area penalty compared to conventional DFFs. Our proposed DFF circuit meets the above requirements.

6.2.1 DFF Circuit Design and Near-/Sub-threshold Operation

Fig. 6.1 depicts the schematic of the 16-transistor charge-pumped DFF (CPDFF), which adopts the master-slave structure. The CPDFF is controlled by three timing signals — an external clock signal (CLK) and two internal clock signals ($CLKH$ and $CLKL$). $CLKH$ and $CLKL$ are generated by two embedded charge pumps, one positive charge pump and one negative charge pump. The $CLKH$ has a higher voltage above VDD while the $CLKL$ has a lower voltage below GND as Fig. 6.1 illustrates. To minimize area penalty, the two charge pumps are shared by 8 DFFs. This results in 1.75 device count increase for each DFF. Yet it still has the smallest transistor count compared to other static CMOS DFFs.

Fig. 6.1 Schematic of proposed charge-pumped DFF.

Fig. 6.2 Simulated output waveforms of the two charge pumps at 0.2 V, 1 kHz.

The proposed DFF, as the conventional TGFF, samples and latches the data at two clock phases. When CLK is low, the master stage samples the input (D) with $CLKL$. Since $CLKL$ has a lower voltage level than GND in this state, it enhances $|V_{gs}|$ of the first pass gate ($PG1$) to ease its state transition, which is particularly beneficial for ultra-low voltage operation. Similarly, when CLK is high, the positively boosted $CLKH$ improves the sampling ability of the slave stage by expanding the overdrive voltage of $PG2$. This also results in an increased output swing at sub-threshold regime.

The logic size is optimized to improve the performance of the DFF. Four inverters with different sizes ($Inv1$, $Inv2$, $Inv3$ and $Inv4$ in Fig. 6.1) are applied in the data propagation path and the positive feedback paths. $Inv1$ is maximally sized to sharpen the data transition and reduce the short-circuit current of $PG1$ during switching. $Inv3$ with the minimum device width is deployed in the feedback loops to reduce the capacitive load of $Inv2$ and decrease the overpower effect from previous storage. The device sizes of $Inv2$ and $Inv4$ are tuned for the best C-Q delay.

The two charge pumps generate a voltage less than GND and a voltage greater than VDD, respectively. The capacitor $C1$ charges its negative terminal to the most

negative voltage (*GND_low*) while *C2* charges its positive terminal to the most positive voltage (*VDD_high*). The pumped voltages are transferred to feed *PG1* and *PG2*. Both capacitors are implemented by metal layers which are laid above the transistors in layout to minimize area penalty. Each charge pump leverages a diode connected NMOS transistor to clamp the pumped voltage. Fig. 6.2 shows the boosted outputs from the two charge pumps at 0.4 V with a clock frequency of 1 kHz.

6.2.2 Inverse-Narrow-Width-Effect-Aware Sizing Strategy

Fig. 6.3 NMOS threshold voltage vs. transistor width at different supply voltages with a 90 nm CMOS technology [61].

Research in [61] has revealed that the Inverse-Narrow-Width Effect (INWE) significantly influences threshold voltage and the corresponding drain current in the near- and sub-threshold regions. The INWE is caused by the parasitic transistor at the sharp corner in the shallow-trench isolation (STI) process. The parasitic transistor will be switched on at lower voltages than the main channel due to the geometry effect of the STI corner. As the transistor width shrinks, it dominates the performance of the whole transistor and makes the threshold voltage lower for narrower transistors. Fig. 6.3 investigates the impact of INWE on the threshold voltage at different voltages for a 90 nm CMOS technology. As the transistor width becomes less than 0.5 μm , the threshold voltage decreases quickly. The variation of the threshold voltage is around 130 mV as the width increases from the minimum value to 0.5 μm [61].

At sub-threshold region, the drain current is inversely proportional to the threshold voltage. As the threshold voltage is lowered, the drain current increases following Equation (2.1). This can be leveraged to enhance the performance of a narrow-

width transistor especially at ultra-low voltage. To utilize it, the total width of a transistor is implemented as a combination of multiple fingered minimum widths [61]. Consequently, each minimum-width finger has the lowest threshold voltage to maximize the drain current. In addition, the drain current of the transistor is accordingly proportional to the width. The C-Q delay is consequently improved with the sizing strategy, which is extremely advantageous for sub-threshold operation. The detailed analysis of C-Q delay improvement is presented in Section 6.3.1.

6.3 Analysis of CPDFF with TGFF and ACFF

The proposed CPDFF is compared with the TGFF and the ACFF through simulations on C-Q delay, setup time, hold time and energy-delay product. The schematics of TGFF and ACFF are shown in Fig.2.13 and Fig. 2.16, respectively. The transistor sizes of the two DFFs are optimized to achieve the best C-Q delay through extensive simulations for fair comparison. The data input and the clock signal fed to 8 CPDFFs, a TGFF and an ACFF come from two-stage buffers. To mimic practical scenarios, a FO4 buffer loaded with 3 pF capacitor is connected to each DFF output.

6.3.1 C-Q Delay Investigation

The utilization of INWE-aware sizing strategy can boost C-Q delay. Simulated delay savings from our proposed CPDFF are listed in Table 6.1. As output Q is logic ‘1’, more than 14% faster C-Q delay is obtained at $V_{DD} = 0.4$ V. When input data is logic ‘0’, the CPDFF can speed up the C-Q delay by 4%. More significant enhancement can be attained if the $PG2$ size is increased. The performance boosting is also observed at lower supply voltage.

In nanometer technologies, DFF performance variation due to transistor mismatch is challenging. To investigate it, a Monte Carlo simulation with 3σ mismatch is conducted. Fig. 6.4 captures the output waveforms of the 3 DFFs at $V_{DD} = 0.4$ V. The TGFF and the CPDFF exhibit better mismatch tolerance than the ACFF, whose output can have a variation of 0.75 μ s when $T = 2$ μ s as Fig. 6.4 shows.

Table 6.1 Performance improvement from INWE-aware sizing strategy.

VDD (V)	T (ns)	Δt of data ‘1’ (ns)	$\Delta t/T$	Δt of data ‘0’ (ns)	$\Delta t/T$
0.7	20	2.2	11%	0.2	1%
0.6	40	10.4	26%	0.8	2%
0.5	300	34	11.3%	4.7	1.6%
0.4	1000	149.3	14.9%	39.8	4%

Fig. 6.4 Simulated output waveforms of CPDFF, TGFF and ACFF at 0.4 V.

Occasionally, it fails in functionality when the CLK triggers. This is due to the extremely reduced output swings of the pass gates at ultra-low voltage despite the aid of adaptive coupling transistors.

Further analysis discloses that the proposed DFF has better performance over the TGFF. This is because the increased $|V_{gs}|$ by charge pumps not only eases state transition of a pass gate but ameliorates its variability as well. A 1000-point Monte

(a)

(b)

Fig. 6.5 Monte Carlo simulation results of C-Q delay: (a) data '0' and (b) data '1'. The proposed CPDFF shows less variability.

Carlo simulation (Fig. 6.5) verifies that the CPDFF can provide a smaller mean value of C-Q delay with less variation than the TGFF.

6.3.2 Comparison of Setup Time and Hold Time

(a)

(b)

Fig. 6.6(a) Setup time of CPDFF, TGFF and ACFF at different process corners and (b) Hold time of CPDFF, TGFF and ACFF at different process corners.

As discussed in Section 2.3.1, T_{setup} and T_{hold} are key figures of merit for DFF circuits. Setup time (T_{setup}) violation can cause input data sampling failure and the consequently functional failure of the DFF. It can be ameliorated by relaxing the clock frequency of the system. Hold time (T_{hold}) violation also causes harsh

functional problem. However, it cannot be compensated by clock frequency manipulation. Therefore, hold time violation is more severe than setup time violation in terms of lack of rectification methods. The master-slave type of DFFs, such as the TGFF and the proposed CPDFF, usually has positive setup time and negative hold time. The negative hold time originates from the fact that the preliminary data is latched by the master stage. It relaxes the requirement that the data should remain unchanged after the clock edge. The negative hold time with the positive setup time makes the master-slave style DFFs not prone to data race [11].

Swept T_{setup} and T_{hold} against VDD are illustrated in Fig. 6.6(a) and (b), respectively. At TT corner, the CPDFF has a moderate setup time whereas the ACFF needs a maximum setup time. At SS corner and VDD = 0.4 V, the ACFF fails due to the extremely degraded output swing of pass gates. Therefore, its setup time cannot be obtained under this circumstance. Similarly, at SS corner, the hold time of the ACFF is not attainable (Fig. 6.6(b)). It reveals that the ACFF circuit is prone to fail at ultra-low voltage and skew process condition. Negative hold time with large absolute value is preferred. The CPDFF, as Fig. 6.6(b) indicates, provides a small negative hold time but generates less variability than the TGFF does.

6.3.3 Analysis of Energy-Delay Product

Fig. 6.7 Simulated Energy-Delay product against data activity at $V_{DD} = 0.4$ V.

Energy-delay space analysis is an effective means to compare the utility of various DFFs [11],[62]. As ultra-low voltage/power applications emerge, an in-depth understanding of energy-delay (ED) tradeoff is crucial to fairly evaluate both energy and performance. A wide range of different ED tradeoffs can be explored by varying the components i and j in the figure of merit of $E^i D^j$. The investigation of minimum $ED^2 \sim ED^5$ is a high-performance-emphasis approach while the exploration of minimum $E^2 D$ and $E^3 D$ is more low power biased. However, the basic energy-delay product is adequate to equally weight energy and delay to examine the features at ultra-low voltage domain.

Fig. 6.7 plots the curves of energy-delay products from the CPDFF, the TGFF and the ACFF with respective to data activity α . When α is 0%, ACFF is still the most efficient DFF style. The CPDFF is not the best if $\alpha < 40\%$ as Fig. 6.7 presents. Because when α is low or even 0%, charge pumps still consume active power, which degrades the energy-delay efficiency. However, when α increases, the advantage of utilizing the CPDFF instead of the ACFF becomes more obvious. At 100% data activity, the CPDFF exhibits more than 30% improvement in energy-delay compared to the ACFF. In addition, the CPDFF is always more efficient than the TGFF regardless of data activity. With respect to power, it follows a similar trend and the experimental data will be analyzed in Section 6.4.

6.4 Test Chip Implementation and Measurement

Synchronous FIFO

Fig. 6.8 Architecture of the FIFO circuit.

Fig. 6.9 Measured C-Q delay against VDD.

A test chip comprising 9 DFFs, 2 charge pumps and 2 FIFOs (First-In-First-Out) is fabricated in a 180 nm technology with a nominal voltage of 1.8 V. One TGFF and eight CPDFF circuits sharing the two charge pumps are implemented to test timing parameters, power consumption and energy-delay product. The outputs of the DFFs are buffered with 2 stages of FO4 inverters, whose power are not incorporated in DFF power calculation. For the two charge pumps, the plate capacitors in the

circuits utilize top two metal layers. By laid over lower layers, area overhead of the capacitors is reduced by 50%. Based on the preliminary simulations, the ACFE circuit is not competent to operate at sub-threshold region due to the functionality issue. Therefore, this circuit is not fabricated. Two 256-bit FIFO circuits are implemented (Fig. 6.8), one deploys the CPDFF and the other utilizes the TGFF. The FIFOs are synthesized with the same control logic and IO circuits. Fig. 6.9 shows the measurement results of C-Q delay against VDD. The CPDFF is fully

(a)

(b)

Fig. 6.10(a) Measured power against VDD. (b) Measured power against frequency.

functional down to 0.18 V with a maximum frequency of 1 kHz. At the minimum voltage, 24.5% delay reduction is observed by the proposed DFF. From 0.3 V to 0.18 V, the CPDFF provides 23% faster C-Q delays than the TGFF on average. Fig. 6.10(a) demonstrate the measurement results of the power against VDD with a frequency of 1 kHz while Fig. 6.10(b) shows the measured power against frequency when VDD = 0.5 V. The power contour in Fig. 6.10(a) shows the CPDFF consumes less power than the TGFF in the whole VDD range, achieving a maximum reduction of 42.3% at 0.5 V and 15.6% at 0.18 V. In Fig. 6.10(b), when sweeping the frequency from 1 kHz to 1 MHz, the power consumption of the CPDFF shoots up to 1.4 $\times$ and the power of the TGFF also augments to 1.2 $\times$. The power reduction by the CPDFF at each operating frequency is depicted with the hollow-dotted curve in the same figure. The CPDFF is more power efficient at low frequency, such as 1 kHz and 10 kHz, and it dissipates 34.1% less at 1 MHz. Fig. 6.11 exhibits the energy-delay products of the two DFF design with respect to data activity α . When VDD scales to 0.5 V, the CPDFF achieves an energy-delay of 11 pJ $\cdot$ ns at 0% data activity, which is 50.9% less than that of the TGFF. As data activity increases, the energy-delay curve of the CPDFF rises slightly and attains 13.1 pJ $\cdot$ ns at $\alpha = 100\%$ whereas the parameter of the TGFF is more than twice. Averagely, the CPDFF

Fig. 6.11 Measured energy-delay product.

suppresses 50.8% power dissipation compared to the TGFF at near-/sub-threshold region. Power reduction of the FIFO by using CPDFF is illustrated in Fig. 6.12. 45% power saving (not including IO and control logic power) is achieved at 0.3 V and 10% data activity thanks to the CPDFF. 31.2% total power is suppressed. Fig. 6.13(a) captures the output waveforms of the CPDFF at the minimum operating voltage. The die micro-photograph is presented in Fig. 6.13(b).

Fig. 6.12 Measured power of the 2 FIFOs at 0.3 V with 10% data activity.

Fig. 6.13(a) Screen capture of CPDFF output waveforms at 0.18 V and (b) die micro-photograph.

6.5 Summary

This chapter presents a 0.18 V energy-delay efficient 16-transistor CPDFF targeting near-/sub-threshold operation. With the aid of charge pumps and INWE-aware sizing strategy, 23% boosted C-Q delay from 0.18 V to 0.3 V is observed. The delay variability is minimized by charge-pumped overdrive voltages. The CPDFF proves to have a 50.8% lower energy-delay product compared to the TGFF. The utility of the CPDFF is verified by a 256-bit FIFO and achieves 31.2% power reduction at 0.3 V. Experimental results validates the proposed CPDFF is competent for near-/sub-threshold applications.

Chapter 7 Conclusions and Future Works

7.1 Conclusions

The research work begins with an investigation of the impact of MTCMOS device on SRAM energy efficiency. Comprehensive simulations reveal the device combinations cause large variations on energy efficiency. Combined with assisting techniques, such as column-interleaved scheme and boosted wordline, the energy efficiency of MTCMOS SRAM can be enhanced as much as 33×.

Leakage and energy efficiency are primary concerns for ultra-low voltage SRAM design. The thesis presents several circuit techniques to implement an energy efficient single-port SRAM with reliable read operation at ultra-low voltage. The proposed 9T SRAM cell with equalized bitline leakage fosters read operation at sub-threshold regime. To further reduce the static energy, MTCMOS technology is utilized to reduce the leakage in the SRAM array. The corresponding degraded energy efficiency is compensated by a CAM-assisted write performance boosting circuit which speeds up the clock frequency.

Disturb due to common-row access is a paramount challenge for dual-port SRAM circuits. The research work explores design techniques to tackle the issue by proposing a 12T dual-port SRAM cell with hierarchical bitline and virtual ground schemes. The novel SRAM cell decreases the probability of suffering disturb and suppresses read disturb at the common-row access condition. The hierarchical write bitlines boost write operation and improve write ability which is originally degraded by write disturb. Test chip has validated a successful common-row access at 0.4 V.

Finally, a 0.18 V energy-delay efficient charge-pumped DFF targeting near-/sub-threshold operation is presented in the research work. The C-Q delay is boosted with the aid of charge pumps and the inverse-narrow-width-effect-aware sizing strategy. The according enhanced overdrive voltage minimizes the delay variability at ultra-low voltage. The proposed DFF proves to have a much lower energy-delay product and be able to work with a supply voltage of 0.18 V.

7.2 Future Works

The ultimate goal of the research work is to develop sub-threshold circuit design techniques for microwatt applications with robustness and high energy efficiency. The research consists of following major tasks covering from design methodology to most essential circuit blocks for microwatt systems: 1) design and optimization techniques for ultra-low voltage digital and memory circuits; 2) design of energy efficient near-/sub-threshold memories; 3) design of energy-delay efficient sub-threshold digital logics, and 4) design of a sub-threshold biomedical signal processor utilizing the circuits proposed by 2) and 3). The thesis has presented energy efficient sub-threshold memories and logic design except the work of biomedical signal processor, which will be presented in the future.

The target application of the biomedical signal processor is a wireless neural SoC platform which has multiple channels with various digital and mixed circuits. Most of the state-of-the-art works use nominal supply voltage in memory domain and logic control domain, which is not efficient with respect to energy consumption. As energy is a topmost design constraint in wireless systems, circuit design techniques for high energy efficiency have to be continuously pursued and explored in the future.

In addition, traditional design methods using HDL coding is not suitable for sub-threshold operation due to the lack of information in standard cell libraries and huge variations. As the ultra-low voltage signal processor with reliable sub-threshold operation and high energy efficiency is highly demanded, novel architectures and circuit techniques for ultra-low power consumption and robustness will be investigated. A standard cell library which is exclusively optimized for sub-threshold operation will be created based on the work from 1) and 3). With the proposed energy efficient SRAMs, it can greatly enhance the performance and power of the processor in future.

Publications

Journal

- [1] **B. Wang**, T. Q. Nguyen, A. Do, J. Zhou. M. Je, and T. Kim, "Design of an ultra-low voltage 9T SRAM with equalized bitline leakage and CAM-assisted energy efficiency improvement," *IEEE Transaction on Circuits and Systems-I (TCAS-I)*, vol. 62, no. 2, pp. 441-448, 2015.
- [2] **B. Wang**, J. Zhou, and T. Kim, "SRAM devices and circuits optimization toward energy efficiency in multi- V_{th} CMOS," *Elsevier Microelectronics Journal (MEJ)*, vol. 46, no. 3, pp. 265-272, 2015.
- [3] X. Liu, J. Zhou, Y. Yang, **B. Wang**, J. Lan, C. Wang, J. Luo, W. Goh, T. Kim, and M. Je, "A 457-nW near-threshold cognitive multi-functional ECG processor CMOS for long-term cardiac monitoring," *IEEE Journal of Solid-State Circuits (JSSC)*, vol. 49, no. 11, pp. 2422-2434, 2014.

Conference

- [4] **B. Wang**, J. Zhou, and T. Kim, "Ultra-low Power 12T Dual Port SRAM for Hardware Accelerators," *IEEE International SoC Design Conference (ISOCC)*, pp. 274-275, Nov. 2014.
- [5] A. Do, Z. Lee, **B. Wang**, I. Chang, and T. Kim, "0.2V 8T SRAM with Improved Bitline Sensing Using Column-based Data Randomization," *IEEE Asian Solid-State Circuits Conference (A-SSCC)*, pp. 141-144, Nov. 2014.
- [6] J. Zhou, X. Liu, C. Wang, K. Chang, J. Luo, J. Lan, L. Liao, Y. Lam, Y. Yang, **B. Wang**, X. Zhang, W. Goh, T. Kim, and M. Je, "A 0.5 V 29 pJ/Cycle Sensor Node Processor for Intelligent Sensing Applications," *IEEE International SoC Design Conference (ISOCC)*, pp. 70-71, Nov. 2014.
- [7] **B. Wang**, J. Zhou, K. H. Chang, M. Je, and T. Kim, "A 0.18V charge-pumped DFF with 50.8% energy-delay reduction for near-/sub-threshold circuits," *IEEE Asian Solid-State Circuits Conference (A-SSCC)*, pp. 121-124, Nov. 2013.

- [8] X. Liu, J. Zhou, Y. Yang, **B. Wang**, J. Lan, C. Wang, J. Luo, W. Goh, T. Kim, and M. Je, "A 457-nW Cognitive Multi-Functional ECG Processor," *IEEE Asian Solid-State Circuits Conference (A-SSCC)*, pp. 141-144, Nov. 2013.
- [9] Y. Yeoh, **B. Wang**, X Yu, and T. Kim, "A 0.4 V 7T SRAM with Write Through Virtual Ground and Ultra-fine Grain Power Gating Switches," *IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 3030-3033, May 2013.
- [10] **B. Wang**, T. Q. Nguyen, A. Do, J. Zhou. M. Je, and T. Kim, "A 0.2V 16Kb 9T SRAM with bitline leakage equalization and CAM-assisted write performance boosting for improving energy efficiency," *IEEE Asian Solid-State Circuits Conference (A-SSCC)*, pp. 73-76, Nov. 2012.
- [11] T. Kim, **B. Wang**, and A. Do, "High Energy Efficient Ultra-low Voltage SRAM Design: Device, Circuit, and Architecture," *International SoC Design Conference (ISOCC)*, pp. 367-370, Nov. 2012.
- [12] Q. Li, **B. Wang**, and T. Kim, "A 5.61 pJ, 16 kb 9T SRAM with Single-ended Equalized Bitlines and Fast Local Write-back for Cell Stability Improvement," *IEEE European Solid-State Device Research Conference (ESSDERC)*, pp. 201-204, Sep. 2012.
- [13] **B. Wang**, J. Zhou, and T. Kim, "Maximization of SRAM Energy Efficiency Utilizing MTCMOS Technology," *Asia Symposium on Quality Electronic Design (ASQED)*, pp. 35-40, Jul. 2012.

Bibliography

- [1] M. Bohr, "The new era of scaling in an SoC world," *ISSCC Dig. Tech. Papers*, pp. 23-28, 2009.
- [2] S. Damaraju, V. George, S. Jahagirdar, T. Khondker, R. Milstrey, S. Sarkar, S. Scott, I. Stolerio, and A. Subbiah, "A 22nm IA multi-CPU and GPU System-on-Chip," *ISSCC Dig. Tech. Papers*, pp. 56-57, 2012.
- [3] R. Islam, A. Sabbavarapu, and R. Patel, "Power Reduction Schemes in Next Generation Intel ATOM Processor based SoC for Handheld Applications," *IEEE Symposium on VLSI Circuits*, pp. 173-174, 2010.
- [4] H. Lakdawala, M. Schaecher, C. Fu, R. Limaye, J. Duster, Y. Tan, A. Balankutty, E. Alpman, C. Lee, S. Suzuki, B. Carlton, H. Kim, M. Verhelst, S. Pellerano, T. Kim, D. Srivastava, S. Venkatesan, H. Lee, P. Vandervoorn, J. Rizk, C. Jan, K. Soumyanath, and S. Ramamurthy, "32nm x86 OS-Compliant PC On-Chip with Dual-Core Atom Processor and RF WiFi Transceiver," *ISSCC Dig. Tech. Papers*, pp. 62-64, 2012.
- [5] S. Rusu, "Microprocessor Design in the Nanoscale Era," http://www.ieee-jp.org/section/kansai/chapter/sscs/20120719/uP_Design_July_2012.pdf
- [6] Q. Li, B. Wang, and T. Kim, "A 5.61 pJ, 16 kb 9T SRAM with Single-ended Equalized Bitlines and Fast Local Write-back for Cell Stability Improvement," *Proc. of the European Solid-State Device Research Conference*, pp. 201-204, 2012.
- [7] A. Wang and A. Chandrakasan, "A 180-mV sub-threshold FFT processor using a minimum energy design methodology," *IEEE J. Solid-State Circuits*, vol. 40, no. 1, pp. 310–319, 2005.
- [8] S. Hanson et al., "A Low-Voltage Processor for Sensing Applications with Picowatt Standby Mode," *IEEE J. Solid-State Circuits*, vol. 44, no. 4, pp. 1145-1155, 2009.
- [9] G. Chen et al., "Millimeter-Scale Nearly Perpetual Sensor System with Stacked Battery and Solar Cells," *ISSCC Dig. Tech. Papers*, pp. 288-289, 2010.

- [10] K. Lee and N. Verma, "A 1.2-0.55V General-purpose Biomedical Processor with Configurable Machine-learning Accelerators for High-order Patient-adaptive Monitoring," *Proc. Eur. Solid-State Circuits Conf. (ESSCIRC)*, pp. 285-288, 2012.
- [11] M. Alioto, E. Consoli, and G. Palumbo, "Analysis and Comparison in the Energy-Delay-Area Domain of Nanometer CMOS Flip-Flops: Part I — Methodology and Design Strategies," *IEEE Trans. on VLSI Systems*, pp. 725-736, 2011.
- [12] M. Alioto, E. Consoli, and G. Palumbo, "Analysis and Comparison in the Energy-Delay-Area Domain of Nanometer CMOS Flip-Flops: Part II — Results and Figures of Merit," *IEEE Trans. on VLSI Systems*, pp. 737-750, 2011.
- [13] L. Chang, Y. Nakamura, R. Montoye, J. Sawada, A. Martin, K. Kinoshita, F. Gebara, K. Agarwal, D. Acharyya, W. Haensch, K. Hosokawa, and D. Jamsek, "A 5.3GHz 8T-SRAM with Operation Down to 0.41 V in 65 nm CMOS," *IEEE Symposium on VLSI Circuits*, pp. 250-253, 2007.
- [14] Y. Wang, U. Bhattacharya, F. Hamzaoglu, P. Kolar, Y. Ng, L. Wei, Y. Zhang, K. Zhang, and M. Bohr, "A 4.0 GHz 291 Mb Voltage-scalable SRAM Design in a 32 nm High-k+ Metal-gate CMOS Technology With Integrated Power Management," *IEEE J. Solid-State Circuits*, vol. 45, no. 1, pp. 103-110, 2010.
- [15] P. Kolar, E. Karl, U. Bhattacharya, F. Hamzaoglu, H. Nho, Y. Ng, Y. Wang, and K. Zhang, "A 32 nm High-k Metal Gate SRAM With Adaptive Dynamic Stability Enhancement for Low-Voltage," *IEEE J. Solid-State Circuits*, vol. 46, no. 1, pp. 76-84, 2011.
- [16] T. Kim, J. Liu, and C. Kim, "A Voltage Scalable 0.26 V, 64 kb 8T SRAM with Vmin Lowering Techniques and Deep Sleep Mode," *IEEE J. Solid-State Circuits*, vol. 44, no. 6, pp. 1785-1795, 2009.
- [17] B. Calhoun and A. Chandrakasan, "A 256-kb 65-nm sub-threshold SRAM Design for Ultra-Low-Voltage Operation," *IEEE J. Solid-State Circuits*, vol. 42, no. 3, pp. 680-688, 2007.

- [18] T. Kim, J. Liu, J. Keane, and C. Kim, "A 0.2 V, 480 kb Subthreshold SRAM With 1k Cells Per Bitline for Ultra-Low-Voltage Computing," *IEEE J. Solid-State Circuits*, vol. 43, no. 2, pp. 518-529, 2008.
- [19] T. Song, W. Rim, J. Jung, G. Yang, J. Park, S. Park, K. Baek, S. Baek, S. Oh, J. Jung, S. Kim, G. Kim, J. Kim, Y. Lee, K. Kim, S. Sim, J. Yoon, and K. Choi, "A 14nm FinFET 128Mb 6T SRAM with V_{MIN} -Enhancement Techniques for Low-Power Applications," *ISSCC Dig. Tech. Papers*, pp. 232-232, 2014.
- [20] Y. Chen, W. Chan, W. Wu, H. Liao, K. Pan, J. Liaw, T. Chung, Q. Li, G. Chang, C. Lin, M. Chiang, S. Wu, S. Natarajan, and J. Chang, "A 16nm 128Mb SRAM in High-k Metal-Gate FinFET Technology with Write-Assist Circuitry for Low- V_{MIN} Applications," *ISSCC Dig. Tech. Papers*, pp. 238-239, 2014.
- [21] J. Chang, Y. Chen, H. Cheng, W. Chan, H. Liao, Q. Li, S. Chang, S. Natarajan, R. Lee, P. Wang, S. Lin, C. Wu, K. Cheng, M. Cao, and G. Chang, "A 20nm 112Mb SRAM in High-k Metal-Gate with Assist Circuitry for Low-Leakage and Low- V_{MIN} Applications," *ISSCC Dig. Tech. Papers*, pp. 316-318, 2013.
- [22] K. Agawa, H. Hara, T. Takayanagi, and T. Kuroda, "A Bitline Leakage Compensation Scheme for Low-Voltage SRAMs," *IEEE J. Solid-State Circuits*, vol. 36, no. 5, pp. 726-734, 2001.
- [23] A. Calimera, A. Macii, E. Macii, and M. Poncino, "Design Techniques and Architectures for Low-Leakage SRAMs," *IEEE Trans. Circuits and Systems — I*, vol. 59, no. 9, pp. 1992-2007, 2012.
- [24] Y. Lai and S. Huang, "X-Calibration: A Technique for Combating Excessive Bitline Leakage Current in Nanometer SRAM Designs," *IEEE J. Solid-State Circuits*, vol. 43, no. 9, pp. 1964-1971, 2008.
- [25] C. Lo and S. Huang, "P-P-N Based 10T SRAM Cell for Low-Leakage and Resilient Subthreshold Operation," *IEEE J. Solid-State Circuits*, vol. 46, no. 3, pp. 695-704, 2011.

- [26] J. Kim, Y. Choi, J. Jeong, S. Lee, and S. Kim, "The v2.0+ EDR Blue-tooth SoC Architecture for Multimedia," *IEEE Trans. Consum. Electron.*, vol. 52, no. 2, pp. 436-444, 2006.
- [27] T. Shiota, K. Kawasaki, Y. Kawabe, W. Shibamoto, A. Sato, T. Hashimoto, F. Hayakawa, S. Tago, H. Okano, Y. Nakamura, H. Miyake, A. Suga, and H. takahashi, "A 51.2 GOPS 1.0 GB/s-DMA Single-chip Multi-processor Integrating Quadruple 8-way VLIW processors," *ISSCC Dig. Tech. Papers*, pp. 194-195, 2005.
- [28] M. Makajima, T. Yamamoto, M. Yamasaki, K. Kaneko, and T. Hosoki, "Homogeneous Dual-processor Core with Shared L1 Cache for Mobile Multimedia SoC," *IEEE Symposium on VLSI Circuits*, pp. 216-217, 2007.
- [29] M. Miyama, J. Miyakoshi, Y. Kuroda, K. Imamura, H. Hashimoto, and M. Yoshimoto, "A sub-mW MPEG-4 Motion Estimation Processor Core for Mobile Video Application," *IEEE J. Solid-State Circuits*, vol. 39, no. 9, pp. 1562-1570, 2004.
- [30] K. Nii, Y. Tsukamoto, M. Yabuuchi, Y. Masuda, S. Imaoka, K. Usui, S. Ohbyashi, H. Makino, and H. Shinohara, "Synchronous Ultra-High-Density 2RW Dual-Port 8T-SRAM with Circumvention of Simultaneous Common-Row-Access," *IEEE J. Solid-State Circuits*, vol. 44, no. 3, pp. 977-986, 2009.
- [31] M. Yabuuchi, Y. Tsukamoto, M. Morimoto, M. Tanaka, and K. Nii, "20nm High-Density Single-Port and Dual-Port SRAMs with Wordline-Voltage-Adjustment System for Read/Write Assists," *ISSCC Dig. Tech. Papers*, pp. 234-235, 2014.
- [32] Y. Ishii, H. Fujiwara, K. Nii, H. Chigasaki, O. Kuromiya, T. Saiki, A. Miyanishi, and Y. Kihara, "A 28-nm Dual-Port SRAM Macro with Active Bitline Equalizing Circuitry against Write Disturb Issue," *IEEE Symposium on VLSI Circuits*, pp. 99-100, 2010.
- [33] W. Dally and J. Poulton, "Digital Systems Engineering," *Cambridge University Press*, pp. 574, 1998.
- [34] R. Baker, "CMOS Circuit Design, Layout, and Simulation," *IEEE Press, Wiley*, pp. 388-389, 2010.

- [35] C. Chang and P. Gupta, "Calibration of Setup and Hold Time for Latches and Flip-Flops," nanocad.ee.ucla.edu/pub/Main/SnippetTutorial/calI.pdf
- [36] C. Chen, K. Bowman, C. Augustine, Z. Zhang, and J. Tschanz, "Minimum Supply Voltage for Sequential Logic Circuits in a 22nm Technology," *IEEE International Symp. Low-Power Electronics and Design*, pp. 181-186, 2013.
- [37] C. Teh, T. Fujita, H. Hara, and M. Hamada, "A 77% Energy-Saving 22-Transistor Single-Phase-Clocking D-Flip-Flop with Adaptive-Coupling Configuration in 40nm CMOS," *ISSCC Dig. Tech. Papers*, pp. 338-340, 2011.
- [38] Y. Kim, W. Jung, I. Lee, Q. Dong, M. Henry, D. Sylvester, and D. Blaauw, "A Static Contention-Free Single-Phase-Clocked 24T Flip-Flop in 45nm for Low-Power Applications," *ISSCC Dig. Tech. Papers*, pp. 466-467, 2014.
- [39] S. Jain, S. Khare, S. Yada, V. Ambili, P. Salihundam, S. Ramani, S. Muthukumar, M. Srinivasan, A. Kumar, S. Gb, R. Ramanarayanan, V. Erraguntla, J. Howard, S. Vangal, S. Dighe, G. Ruhl, P. Aseron, H. Wilson, N. Borkar, V. De, and S. Borkar, "A 280mV-to-1.2V Wide-Operating-Range IA-32 Processor in 32nm CMOS," *ISSCC Dig. Tech. Papers*, pp. 66-68, 2012.
- [40] P. Liu, J. Wang, M. Phan, M. Garg, R. Zhang, A. Cassier, L. Chua-Eoan, B. Andreev, S. Weyland, S. Ekbote, M. Han, J. Fischer, G. Yeap, P. Wang, Q. Li, C. Hou, S. Lee, Y. Wang, S. Lin, M. Cao, and Y. Mii, "A dual core oxide 8T SRAM cell with low Vccmin and dual voltage supplies in 45nm triple gate oxide and multi Vt CMOS for very high performance yet low leakage mobile SoC applications," *IEEE Symposium on VLSI Technology*, pp. 135-136, 2010.
- [41] C. Diaz, K. Young, J. Hsu, J. Lin, C. Hou, C. Lin, J. Liaw, C. Wu, C. Su, C. Wang, J. Ting, S. Yang, K. Lee, S. Wu, C. Tsai, H. Tao, S. Jang, S. Shue, H. Hsieh, Y. Wang, C. Chen, S. Yang, S. Fu, S. Chang, T. Lo, J. Wu, J. Shy, C. Liu, S. Chen, B. Lin, B. Liew, T. Yen, C. Yu, Y. Chao, M. Liang, C. Wang, and J. Sun, "A 0.18 μm CMOS Logic Technology with Dual Gate Oxide and Low-k Interconnect for High-Performance and Low-Power Applications," *IEEE Symposium on VLSI Technology*, pp. 11-12, 1999.

- [42] M. E. Sinangil, N. Verma, and A. Chandrakasan, "A 54 nm 0.5V 8T column-interleaved SRAM with on-chip reference selection loop for sense amplifier," *IEEE Asian Solid-State Circuits Conf. (A-SSCC)*, pp. 225-228, 2009.
- [43] N. Verma, and A. Chandrakasan, "A 256 kb 65 nm 8T Subthreshold SRAM Employing Sense-Amplifier Redundancy," *IEEE J. Solid-State Circuits*, vol. 43, no. 1, pp. 141-149, 2008.
- [44] H. Kim, Y. Kim, J. Oh, and L. Kim, "A Reconfigurable SIMT Processor for Mobile Ray Tracing with Contention Reduction in Shared Memory," *IEEE Trans. on Circuits and System — I*, vol. 60, no. 4, pp. 938-950, 2013.
- [45] M. Ghaed, G. Chen, R. Haque, M. Wieckowski, Y. Kim, G. Kim, Y. Lee, I. Lee, D. Fick, D. Kim, M. Seok, K. Wise, D. Blaauw, and D. Sylvester, "Circuits for a Cubic-Millimeter Energy-Autonomous Wireless Intraocular Pressure Monitor," *IEEE Trans. on Circuits and System — I*, vol. 60, no. 12, pp. 3152-3162, 2013.
- [46] M. Tu, J. Lin, M. Tsai, C. Lu, Y. Lin, M. Wang, H. Huang, K. Lee, W. Shih, S. Jou, and C. Chuang, "A Single-Ended Disturb-Free 9T Subthreshold SRAM With Cross-Point Data-Aware Write Word-Line Structure, Negative Bit-Line, and Adaptive Read Operation Timing Tracing," *IEEE J. Solid-State Circuits*, vol. 47, no. 6, pp. 1469-1482, 2012.
- [47] A. Teman, L. Pergament, O. Cohen, and A. Fish, "A 250 mV 8 kb 40 nm Ultra-Low Power 9T Supply Feedback SRAM (SF-SRAM)," *IEEE J. Solid-State Circuits*, vol. 46, no. 11, pp. 2713-2726, 2011.
- [48] A. Alvandpour, D. Somasekhar, R. Krishnamurthy, V. De, S. Borkar, and C. Svensson, "Bitline Leakage Equalization for Sub-100nm Caches," *Eur. Solid-State Circuits Conf. (ESSCIRC)*, pp. 401-404, 2003.
- [49] D. Jeon, M. Seok, C. Chakrabarti, D. Blaauw, and D. Sylvester, "A Super-Pipelined Energy Efficiency Subthreshold 240 MS/s FFT Core in 65 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 47, no. 1, pp. 23-34, 2012.
- [50] R. Abdallah and N. Shanbhag, "A 14.5 fJ/cycle/k-Gate, 0.33 V ECG Processor in 45 nm CMOS Using Statistical Error Compensation," *IEEE Custom Integr. Circuits Conf. (CICC)*, 2012, pp. 1-4, 2012.

- [51] B. Wang, J. Zhou, and T. Kim, "Maximization of SRAM Energy Efficiency Utilizing MTCMOS Technology," *Asia Symp. Quality Electronic Design (ASQED)*, pp. 35-40, 2011.
- [52] K. Pagiamtzis and A. Sheikholeslami, "Content-Addressable Memory (CAM) Circuits and Architectures: A Tutorial and Survey," *IEEE J. Solid-State Circuits*, vol. 41, no. 3, pp. 712-727, 2006.
- [53] Y. Sinangil, and A. Chandrakasan, "An Embedded Energy Monitoring Circuit for a 128kbit SRAM with Body-biased Sense-Amplifiers," *Asian Solid-State Circuits Conf. (A-SSCC)*, pp. 69-72, 2012.
- [54] S. Lütke-meier, T. Jungeblut, H. Berge, S. Aunet, M. Porrmann, and U. Ruckert, "A 65 nm 32 b Subthreshold Processor With 9T Multi-V_t SRAM and Adaptive Supply Voltage Control," *IEEE J. Solid-State Circuits*, vol. 48, no. 1, pp. 8-19, 2013.
- [55] M. Chang, M. Chen, L. Chen, S. Yang, Y. Kuo, J. Wu, H. Su, Y. Chu, W. Wu, T. Yang, and H. Yamauchi, "A Sub-0.3 V Area-Efficient L-Shaped 7T SRAM With Read Bitline Swing Expansion Schemes Based on Boosted Read-Bitline, Asymmetric-V_{TH} Read-Port, and Offset Cell VDD Biasing Techniques," *IEEE J. Solid-State Circuits*, vol. 48, no. 10, pp. 2558-2569, 2013.
- [56] M. Yabuuchi, Y. Tsukamoto, M. Morimoto, M. Tanaka, and K. Nii, "20nm High-Density Single-Port and Dual-Port SRAMs with Wordline-Voltage-Adjustment System for Read/Write Assists," *ISSCC Dig. Tech. Papers*, pp. 234-236, 2014.
- [57] Y. Ishii, H. Fujiwara, S. Tanaka, T. Doguchi, O. Kuromiya, H. Chigasaki, Y. Tsukamoto, and K. Nii, "A 28 nm Dual-Port SRAM Macro with Screening Circuitry Against Write-Read Disturb Failure Issues," *Asian Solid-State Circuits Conf. (A-SSCC)*, pp. 1-4, 2010.
- [58] A. Mishra, A. Raymond, L. Amaru, G. Sarkis, C. Leroux, P. Meinerzhagen, A. Burg, and W. Gross, "A Successive Cancellation Decoder ASIC for a 1024-bit Polar Code in 180nm CMOS," *IEEE Asian Solid-State Circuits Conf. (A-SSCC)*, pp. 205-208, 2012.

- [59] F. Hsiao, A. Tang, D. Yang, M. Pham, and M. Chang, "A 7Gb/s SC-FDE/OFDM MMSE Equalizer for 60GHz Wireless Communications," *IEEE Asian Solid-State Circuits Conf. (A-SSCC)*, pp. 293-296, 2011.
- [60] M. Altaf, J. Tillak, Y. Kifle, and J. Yoo, "A 1.83 μ J/Classification Nonlinear Support-Vector-Machine-Based Patient-Specific Seizure Classification SoC," *ISSCC Dig. Tech. Papers*, pp. 100-102, 2013.
- [61] J. Zhou, S. Jayapal, B. Busze, L. Huang, and J. Stuyt, "A 40 nm Inverse-Narrow-Effect-Aware Sub-Threshold Standard Cell Library," *IEEE Trans. on Circuits and System — I*, vol. 59, pp. 2569-2577, 2012.
- [62] E. Consoli, M. Alioto, G. Palumbo, and J. Rabaey, "Conditional Push-Pull Pulsed Latches with 726fJ·ps Energy-Delay Product in 65nm CMOS," *ISSCC Dig. Tech. Papers*, pp. 482-484, 2012.