Low-power Volatile and Non-volatile Memory Design

by

Qing Dong

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering)
in the University of Michigan
2017

Doctoral Committee:

Professor Dennis Michael Sylvester, Chair
Professor David Blaauw
Professor Branko Kerkez
Professor Zhengya Zhang

Qing Dong
qingdong@umich.edu
ORCID ID: 0000-0002-1380-269X

DEDICATION

To my family & friends…

This dissertation is dedicated to my family and friends that supported, encouraged, and inspired me throughout my education.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Embedded memories play a pivotal role in VLSI systems to support the increasing need of data storage in various applications. With technology scaling, memory cell size gets significantly minimized in order to boost the storage capacity. Over half of area in advanced VLSI systems are occupied by embedded memories, especially 6T SRAM which provides fastest performance than others. However, VDDmin of 6T SRAM doesn't scale well in advanced technologies. As a result, 6T SRAM dominates power consumption in advanced VLSI systems like data centers and IoTs which have growing need for large amount of low-power memories.

This thesis presents several circuit and system solutions to reduce power consumption of different types of embedded memories, varying from volatile SRAM to non-volatile memories such as NOR flash and STT-MRAM.

We first describe a 5T SRAM with one transistor less than conventional 6T SRAM. It not only achieves 7.2% area saving than 6T but also improves read margin by decoupling read/write paths. With single-port read with improved read VDDmin, access energy gets significantly reduced. 4Mb of 5T SRAM is applied to a face-recognition machine-learning accelerator.

Second, a 4+2T SRAM cell that uses the N-well as a write wordline is proposed with 15% area saving than 8T SRAM. Decoupled differential read paths significantly improve read noise margin, achieving 0.25V VDDmin and 4fJ/bit access energy. Moreover, reliable multi-word activation is realized for in-memory-computing and BCAM/TCAM applications.

Third, we present a 1Mb sub-100μW embedded NOR flash for battery-powered miniature sensor-node system. Multiple low-power circuit techniques are applied to the high-voltage generation and delivery system and a margin-doubled cross-sampling current sense amplifier is

proposed. Measurements in a 90nm embedded flash technology show 30× and 22× lower program and erase energy, respectively, compared with a standard flash compiler macro.

Fourth, a low-power STT-MRAM in 28nm technology is described. A single-cap based offset-cancelled sense amplifier is proposed to improve sensing margin, and in-situ self-termination write method is used to save write power. We achieve 2.8ns read access time, and over 30% write power consumption is reduced with the proposed write method.

Finally, we explore the feasibility of applying the emerging non-volatile spintronic memory devices into two analog computing applications: analog-to-digital converter and neural network. The analog-to-digital converter using racetrack nanowire can be 1000× smaller than conventional implementation using CMOS. With compact racetrack converter as the neuron, spin rectified-linear and recurrent neural networks can be realized.

CHAPTER 1. Introduction

There has been an ever growing demanding for very large scale integrated (VLSI) systems in broad applications, including mobile computing, consumer electronics, automobile electronics, high-performance data center, and internet of things (IoT). Embedded memories are key components of all VLSI systems. It is often required to have a large amount of embedded memory to support the increasing need of data storage.

CMOS technology scaling is rendering chips smaller, faster and cheaper. Memory benefits a lot from technology scaling in past years, especially for bitcell density and read/write performance. Among embedded memories, 6T SRAM always plays a critical role in all VLSI systems because of its superior speed and full compatibility with logic process technology. As shown in Figure 1.1 [1], the bitcell size of conventional 6T SRAM changes from 0.999μm^2 in 90nm technology to 0.027 μm^2 in 7nm technology, featuring 37 times density improvement. With bitcell size scaling, the capacity of last-level-cache (LLC) in processors has been increased from Mb to Gb within 10 years.

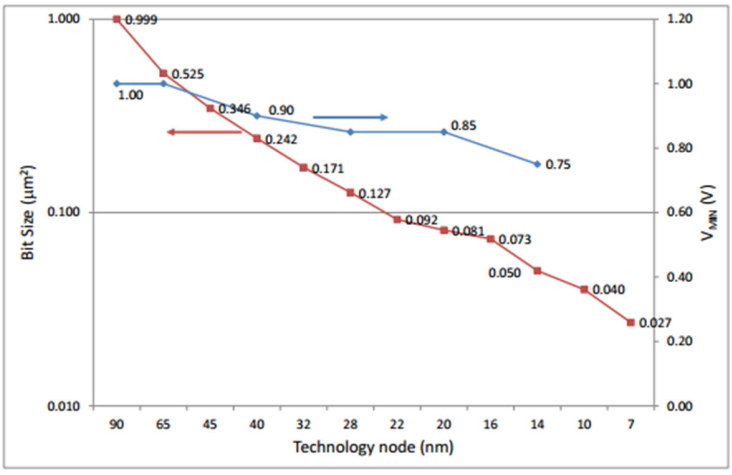


Figure 1.1 SRAM bitcell technology scaling and VDDmin scaling [1].

In most VLSI systems, memory macro dominates the total chip area. As shown in Figure 1.2 [2], more than 50% of chip area is occupied by the L2 and L3 cache, which is consist of 6T SRAM macro, in Intel Xeon processor. 6T SRAM provides highest read/write performance among other types of memories. However, 6T SRAM requires power to be always supplied in order to maintain the data. Due to the required large capacity, on-chip 6T SRAM draws a large amount of leakage current, dominating the system power [2]. Moreover, the power consumption could be even worse when temperature increases due to the exponential temperature dependency of leakage current.

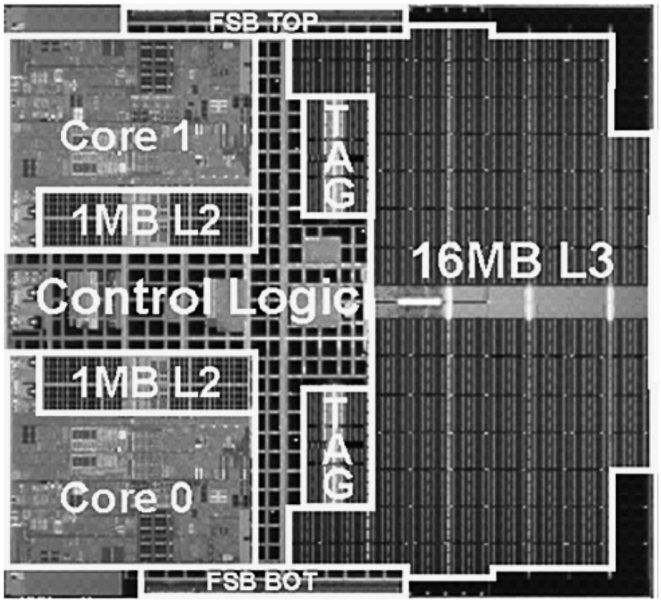


Figure 1.2 Intel Xeon 7100 processor with >50% of area for cache [2].

Low-power consumption is very important for most applications like portable consumer electronics, data centers and IoTs. As for portable consumer electronics, power consumption determines both working time and standby time. Since the battery capacity is always limited, the only method to extend the battery life is to reduce the system power consumption which is largely dominated by memories. With regard to data centers, huge amount of memories are used to meet the massive data storage demanding. US data centers consumed about 70 billion

kilowatt-hours of electricity in 2014. Therefore, energy efficiency threatens the development of data centers. IoTs are emerging applications for memories. Since the typical size of batteries in IoT systems is very small (mm-scale), IoTs have strictest requirement for low-power operation. Conventional memories with mW operation power can drain off the mm-scale battery in seconds. As a result, the power budget of the IoTs could limit the on-chip memory capacity instead of chip area. Therefore, low-power memory is a requisite for VLSI system with limited power budget.

Most common method to reduce the memory power is to lower the supply voltage. However, supply voltage didn't scale that well in advanced technology as shown in Figure 1.1 due to the difficulty of Vth scaling. Moreover, process variation and leakage are becoming more and more severe in advanced technologies, degrading read noise margin of 6T SRAM and limiting its VDDmin scalability. Special optimizations have to be carefully applied to SRAM to solve these problems.

Volatile memories such as SRAM and DRAM use capacitive nodes to store the data, requiring power to be always supplied. Once the power is off, data disappear. To lower standby power, non-volatile memories are introduced which can store data in a permanent way: retaining data even when the power is off. The non-volatile memory cell typically can be switched between stable physical states, which represent binary data. Most recent consumer electronics and automobiles have further broadened the embedded application for non-volatile memories. With increasing demand of non-volatile memory for further scaling of the semiconductor technology, multiple embedded non-volatile memory technologies have been proposed, such as NOR Flash, phase-change RAM (PRAM), ferroelectric RAM (FeRAM), resistive RAM (RRAM) and magnetic RAM (MRAM).

Among various non-volatile memory technologies, embedded NOR flash is most widely-used because of its established process, great CMOS compatibility, small cell size, good yield and reliability. However, NOR flash consumes high power (typically mW) during program and

erase because of its high voltage operation. Therefore, more power-efficient circuit techniques for embedded NOR flash have to be explored to enable their application in VLSI systems with limited power budget.

In recent years, STT-MRAM has attracted a lot of attentions as emerging non-volatile solution, which not only addresses some of the fundamental scaling limits in the conventional NOR flash, but also brings new characteristics to the nonvolatile memories. In addition to random accessing capability, STT-MRAM offers significant improved performance and endurance over embedded NOR flash, which could open up whole new applications like non-volatile LLC, non-volatile processors, and spin-based systems. However, the limited sensing margin still challenges the circuit design for STT-MRAM.

1.1 Low-power SRAMs and In-memory-computing

6T SRAM is the most important memory which has been embedded in almost all of VLSI chips. However, the standby power of 6T SRAM always dominates the system power consumption because of its poor VDDmin scalability limited by degraded read noise margin. Optimizing read noise margin of 6T SRAM at scaled supply voltage demands stronger pull-down transistor or assist techniques like wordline-under-drive [3-5], which induces extra area overhead and power consumption.

There have been many research activities to address this issue with different bit cell structure, such as 7T [6-7], 8T [8-9] and 10T [10-11] SRAM designs. The decoupled access paths in those designs enable optimizing read and write operations independently and hence significantly enhance operation margins. Enlarged margins obtained from such approaches indeed lowers minimum operating voltage of the SRAM, which translates to reduced power consumption. However, those techniques also introduce direct area overhead of bit cell so one must deal with a trade-off between area and voltage scalability. 5T structure has also been

proposed in [12] for area saving, but the margin improvement is limited because of the shared read and write path.

Not only the memory itself, the data transferring also incurs substantial energy and latency costs. As shown in Figure 1.3, conventional Von Neumann architectures involve three major steps: 1) memory reading, 2) registers storing, and 3) ALU computing. They rely on the continual transfer of acquired data between memory and computation elements on-chip, incurring substantial energy and latency costs, dominating the system power consumption and performance. To minimize the energy and performance overhead of excessive memory reads and data transfer, in-memory computing allows for data processing inside the on-chip memories (typically SRAM) [13]. As shown in Figure 1.3, it activates multiple rows simultaneously within the memory subarray itself and computes results directly on the bitline. Results of a computation are available ready immediately after a memory read operation, saving clock cycles and interconnect energy. In this way more efficient computing is enabled by the introduction of more "functional", i.e., one that can perform more than simple reads and writes.

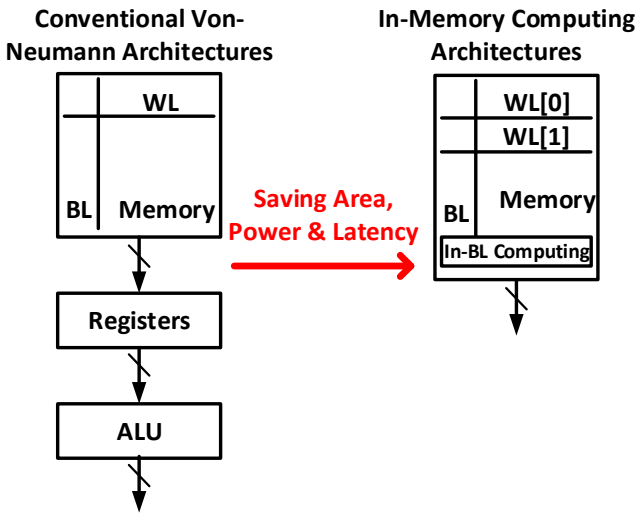


Figure 1.3 Comparison between conventional method and in-memory computing.

1.2 NOR Flash Memory

NOR Flash embeds a floating gate inside the nominal poly gate to absorb electrons. The original state without electrons has low threshold voltage of the floating-gate transistor, representing '1'; while the floating gate with electrons trapped can increase the threshold voltage, representing '0'. Hot carrier injection can move electrons into the floating gate and FN-tunneling helps remove the trapped electrons as shown in Figure 1.4.

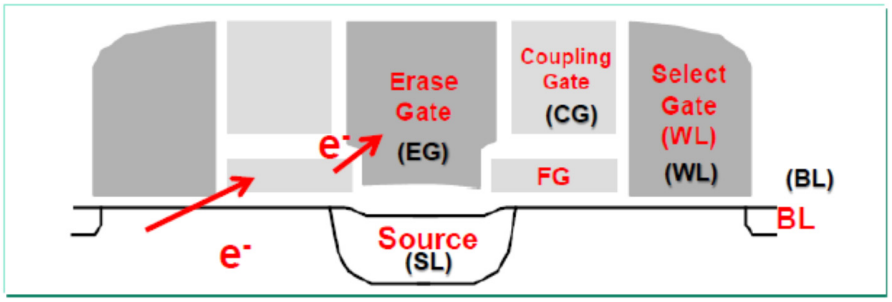


Figure 1.4 Spit-gate NOR Flash Cell.

Compared with 6T SRAM which typically takes ~200F^2 area, NOR flash only occupies ~20F^2 area with 1T or 1.5T structure. The area density gets improved by over 10 times. Moreover, embedded NOR flash memory is also compatible with CMOS technology, making them promising in SOCs requiring non-volatile storage.

As power supply can be fully off during standby, the memory standby power is significantly reduced. This is important for battery-power IoT devices [14-15] as shown in Figure 1.5. They are typically highly duty-cycled system. They will be active for very short time while stay in sleep mode for long stretches. However, conventional NOR flash requires over 10V to perform HCI based program and FN-tunneling based erase operations. It takes mW instantaneous power which drains a mm-scale battery in seconds, limiting its applications in IoT sensor node systems.

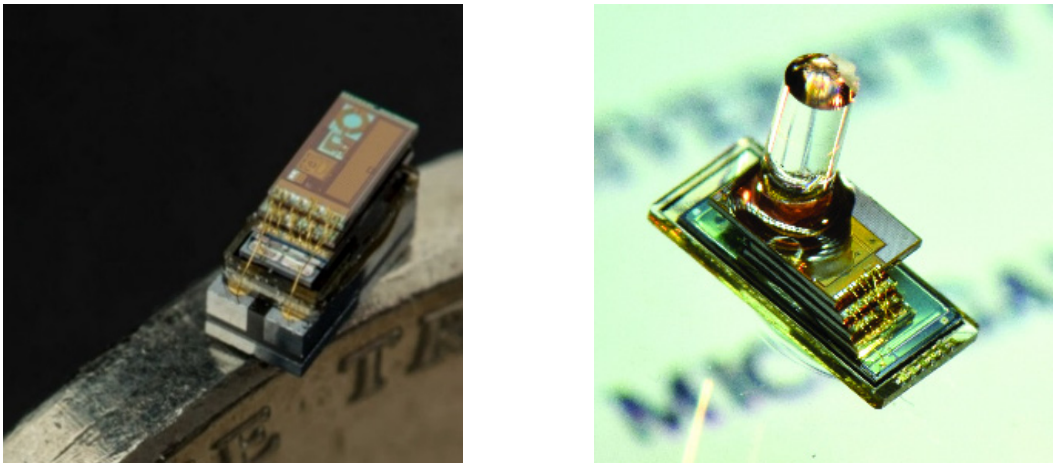


Figure 1.5 Battery-powered mm-scale sensor node systems [14-15].

1.3 STT-MRAM Memory

STT-MRAM uses magnetic tunnel junction (MTJ) to store the data. As shown in Figure 1.6, the magnetic tunnel junction consists of one fixed spin layer, one isolation layer and one free spin layer. If the current flow from free layer to fixed layer, the spin polarization of free layer will become same as that of fixed layer. Then the resistance keeps low, representing '1'. If the current flow in the opposite direction, then the spin polarization of the free layer will be different from that in fixed layer, keeping high resistance ('0').

Compared with NOR flash, STT-MRAM has improved performance, better scalability, higher endurance, and lower write energy. However, the resistance difference between the two states is very small, which can be only twice. As a result, sensing margin is limited, challenging sense amplifier design. Moreover, the required threshold current to flip the state is very high, and therefore, write instantaneous power is quite high.

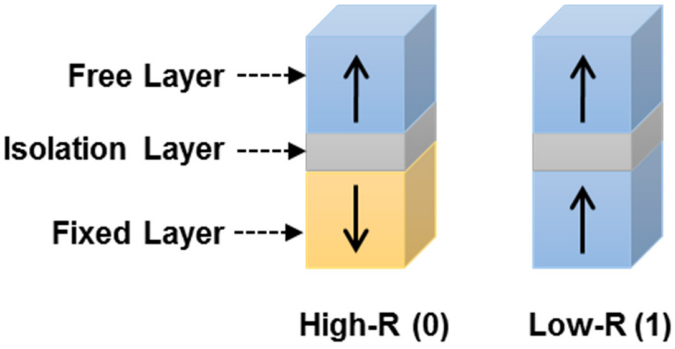


Figure 1.6 Magnetic tunnel junction cell.

1.4 Emerging Spintronic Devices

CMOS meets some challenges with technology scaling such as the leakage and large area. Recently, a number of new materials and novel devices have been proposed to replace CMOS in specific applications. The discovery of current-induced domain wall (DW) motion has driven the invention of several spintronic devices that hold promise for non-volatility, high endurance, high density, and low power [16-17]. With perpendicular magnetic anisotropy (PMA) in CoFeB/MgO structures, multiple magnetic domains separated by DWs can be maintained in one nanowire for multi-bit non-volatile memory [18-22]. Domain wall neurons have also been reported as suitable for current comparison operation and can function as current comparators in SAR ADC [23-24]. Spintronic devices are promising for analog computing and neuromorphic computing.

1.5 Thesis Organization

This dissertation proposes to lower the power consumption of different types of memory, varying from CMOS SRAM to emerging non-volatile memory.

In Chapter 2, this thesis proposes a mostly-read 5T SRAM design, with decoupled read path, which simultaneously provides better read margin, less read access energy and smaller

bit cell size than conventional memory designs such as 6T. The 4Mb 5T SRAM macro is applied to a face-recognition accelerator.

In Chapter 3, this thesis proposes a 4+2T SRAM design that uses the N-well as a write wordline, eliminating the access transistors and resulting in a 4T-core memory cell. Two decoupled read paths (2T) significantly improve read noise margin, enabling reliable multi-word activation for logic operations while limiting area overhead to only 12% over 6T SRAM in the same technology. Using dual sense amplifiers, Boolean logic functions (AND, OR, XOR) between the two activated words can be realized. Furthermore, with separated RBL/RBLB and RWL/RWLB, the SRAM can be configured as a BCAM or TCAM, enabling searching operations.

In Chapter 4, this thesis proposes a 1 Mb embedded NOR Flash memory in 90nm ESF3 technology that is designed for integration into an ultra-low power sensor system. Multiple low-power techniques are applied to minimize the flash write power: 1) combined Dickson and ladder pump topology with MIM caps as flying cap; 2) self-adjusting charge pump regulation loop; 3) Ultra-low power voltage and current reference generation circuits. Also, a cross-sampling current sense amplifier is proposed for sensing margin improvement. The low power NOR flash is incorporated into a complete mm-scale sensor node system to reduce sleep power and extend battery lifetime.

In Chapter 5, this thesis proposes a low-power variation-tolerant 1Mb STT-MRAM in 28nm technology. To improve the read sense margin, a constant current based voltage sensing method using single cap to cancel sense amplifier offset is proposed. Moreover, a variation-tolerant read reference generation method is implemented in the design. To lower the write power, in-situ self-terminated write is used to detect the write end and auto disable the write driver.

In Chapter 6, this thesis proposes one applications using emerging spintronic devices: racetrack ADC. We explored the feasibility of analog-to-digital converter (ADC) based on

current-induced domain wall motion and introduces an n-bit ADC using n racetrack magnetic nanowires. The racetrack ADC is applied to an ultra-high speed digital pixel sensor (DPS) imaging system.

In Chapter 7, this thesis proposes another application using emerging spintronic devices: neural networks. A spin synapse device has been proposed with analog programmability using all charge current. The synapse devices can be placed in a cross-bar array to form a dense neural network. DOT product nano-function can be realized using current summation. With compact racetrack converter as the neuron, spin rectified-linear neural network can be implemented. Storing the DW motion in a time-based fashion, the more complicated RNN can also be realized for time-involved inference tasks.

Finally, in Chapter 8, the conclusion of this dissertation is made by summarizing the proposed circuits and discussing possible future works.

CHAPTER 2. Low-power 5T SRAM

2.1 Introduction

Face recognition can categorize each face and find corresponding person in the database. Figure 2.1 describes a face detection and recognition hardware architecture. This architecture requires more than 4Mb memory in total mostly to save coefficients for algorithm coefficients and hence the memory blocks dominate the system in terms of both area and power consumption. Since most of the system power is dissipated as memory leakage and access energy, optimizing these memory blocks can significantly improve energy efficiency of the entire system. It is observed that more than 90% of the space is required specifically to store algorithm coefficients for PCA and SVM (colored orange in Figure 2.1), which are programmed at the beginning of the face detection and recognition process. Generally no further data update is necessary unless the face database itself is modified or updated, which happens relatively infrequently. Hence typically the memory read operation dominates overall system power consumption. To make use of this observation, a new SRAM design is proposed to primarily optimize low-energy and margin-improved read operation as well as smaller bit cell size.

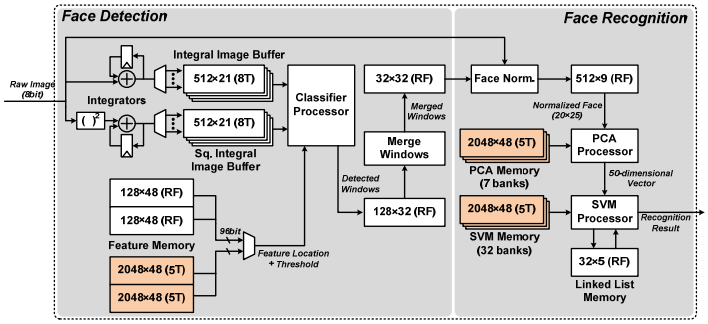


Figure 2.1 Architecture of the face detection and recognition processor.

Due to the required large capacity, the on-chip SRAM storing coefficients draws a large amount of leakage current, dominating the system power [25]. We can effectively suppress the leakages by adopting deep supply voltage scaling [26], but conventional 6T SRAM suffers from degraded read margin under lowered power supply [27]. Decoupled access paths can enable optimizing read and write operations independently and hence significantly enhance operation margins. Enlarged margins obtained from such approaches indeed lowers minimum operating voltage of the SRAM, which translates to reduced power consumption. However, conventional decoupled SRAM like 7T [6-7], 8T [8-9] and 10T [10-11] all introduce considerable area overhead of bit cell so one must deal with a trade-off between area and voltage scalability.

A mostly-read 5T SRAM design is presented with decoupled read path, which simultaneously provides better read margin, less read access energy and smaller bit cell size than conventional memory designs. This work was done as part of a collaborative project on face detection/recognition and that DSP part was done by another student Dongsuk Jeon and my contribution was the entirety of the SRAM design and test.

2.2 Bitcell Design and Decoupled Read

2.2.1 5T SRAM cell

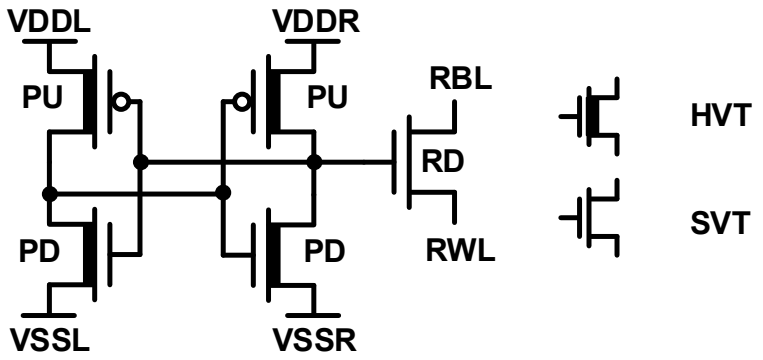


Figure 2.2 Proposed 5T memory bit cell design. VDDL/VDDR and VSSL/VSSR are the left/right power and ground terminals, respectively.

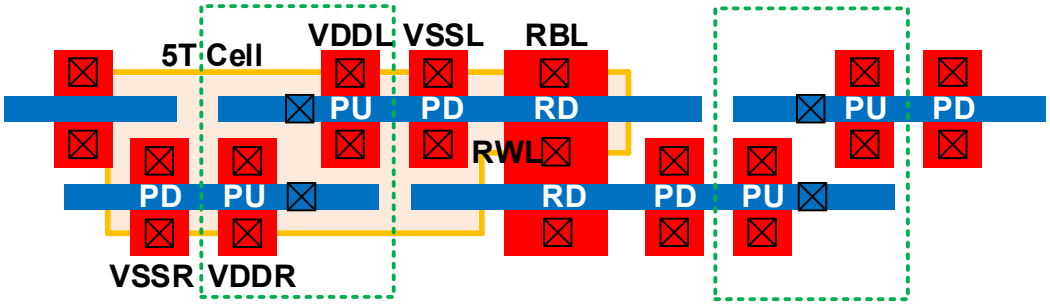


Figure 2.3 Layout of 5T bit cell. Isolated read and write paths allow for minimum-sized pull-up and pull-down devices.

Fig. 2.2 shows the proposed SRAM bit cell. In order to save area, we would like to have the minimum number of transistors in each bit cell. The bit cell has 4 transistors for back-to-back inverters storing data internally and an additional transistor for bit cell access, resulting in 5-transistors structure. Decoupled read access scheme significantly improves read margin and enables separate leakage optimization. According to Monte Carlo simulation, the mean dynamic RNM of 6T SRAM at 0.6V without any read assist is only 0.083V with a standard deviation of 0.019V, whereas that of 5T SRAM is 0.224V with 0.022V standard deviation. If we define the VCCmin in simulation as the power supply at which mean/sigma of RNM is 6, then the simulated VCCmin for 6T and 5T are 0.77V and 0.36V, respectively.

The inverters use minimum-sized HVT devices for leakage reduction, while the SVT access transistor allows for fast and reliable readout. The leakage of the proposed HVT 5T cell is only 19% of conventional 6T which uses SVT to balance read margin and read speed. Figure 2.3 shows the L-shape layout of the proposed 5T bit cell. Similar to 6T and its variants including 7T [6], the read word line is shared with the next bit cell in the same row. To minimize area overhead, VDD and VSS rails are also shared with adjacent rows and columns, respectively, similar to lithographically-symmetric 6T layout. Since it has decoupled read and write paths,

minimum size transistors can be employed for both pull-up and pull-down devices and the proposed cell has 7.2% smaller area than a standard 6T using logic rules.

2.2.2 Decoupled Read

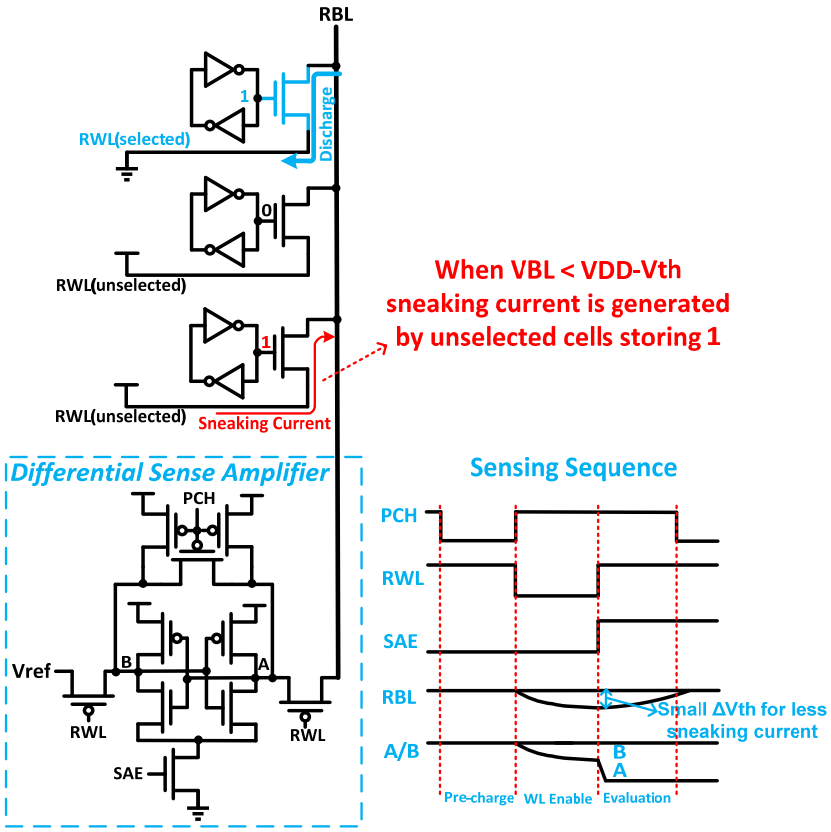


Figure 2.4 Readout path of 5T memory.

Figure 2.4 shows the readout circuitry. A differential cross-coupled sense amplifier is used instead of single-ended inverter to accelerate read speed and reduce access power consumption. During read, selected RWL is driven to ground while unselected RWLs remain high. If the cell stores 0, RBL will stay high; otherwise, RBL will be discharged. Because of the sneaking currents from unselected RWLs, the RBL cannot be fully discharged to ground, which incurs short circuit current if single-ended inverter is used as sensing circuitry. Therefore, differential

cross-coupled sense amplifier is used here so that the readout circuitry can distinguish small voltage difference even before the sneaking currents appears.

Compared with inverter based sensing, the differential cross-coupled sense amplifier can improve read speed by 30% because of small signal sensing and save power by 25% due to elimination of sneaking current and short-circuit current. Although the area is doubled compared to inverter, the area overhead is still less than 1% of the whole macro. From 10k Monte Carlo simulations, the standard deviation of the input offset is 11.2mV and the mean evaluation time is 160ps in 40nm technology. The sensing sequence is detailed in Figure 2.4.

2.3 Write Operation

2.3.1 Basic Write Method

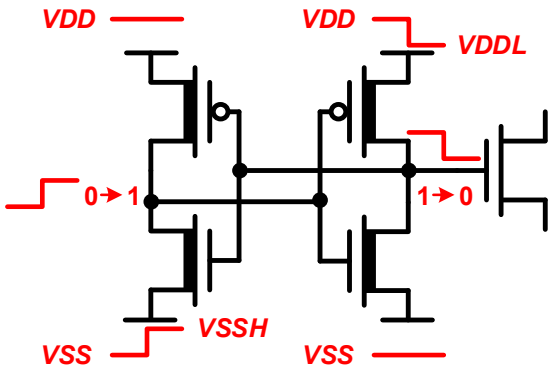


Figure 2.5 Basic write operation of 5T memory. VDDL is the lowered voltage level of VDD and VSSH is the raised voltage level of VSS.

Since the proposed cell has no pass-gate transistor connected to the storage nodes, the power and ground rails are used to write values into the cell by changing their voltages dynamically. Figure 2.5 describes basic write operation in detail. Assume that a bit cell is storing a '1' in the right internal node. Then the left VSS rail is raised to an intermediate voltage VSSH and the left

internal node voltage follows it since the pull-down transistor is on. Similarly, lowering the right VDD rail to another intermediate voltage VDDL also decreases the right internal node voltage. As this process continues, at some point the internal values becomes flipped and the value '0' is successfully written into the cell. The value '1' can be written by changing the opposite VDD and VSS rails. Raising VSS or lowering VDD does not change the cell state, but combining both flips the cell.

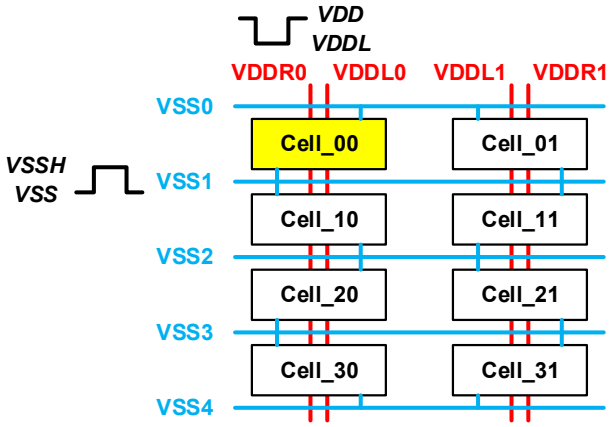


Figure 2.6 Example of write disturbance issue. Writing into Cell_00 also affects other bit cells including Cell_10.

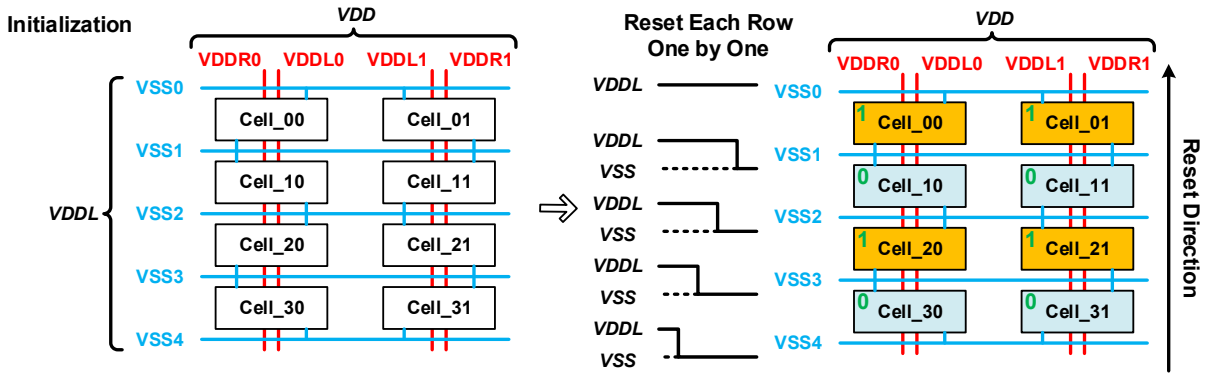


Figure 2.7 Memory reset scheme.

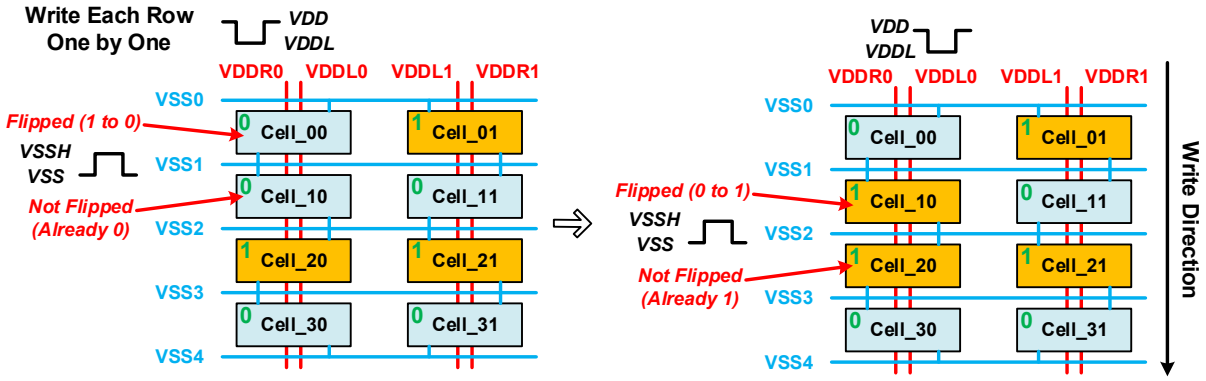


Figure 2.8 Sequential write scheme.

As the compact lithographically-symmetric layout is used, the shared power and ground rails (Figure 2.3) cause one important issue with write operation that need to be considered. Assume we write a value in the highlighted cell (Cell_00) in Figure 2.6. We need to write a value by raising VSS1 rail, but since the rail is shared across the cells in the current and next rows it may also flip the values stored in other cells sharing the same rail. We also need to lower the VDDR0 rail for write operation and it may disturb other data stored in the same column. To avoid the write disturbance issue, a special systematic write scheme is developed. As mentioned earlier, the system memory is programmed at the beginning of recognition process and need to be updated very rarely. Taking advantage of this property, before writing any value into the memory firstly the entire macro will be reset as shown in Figure 2.7. All the VSS rails are initially tied to the intermediate voltage VDDL, and each row is returned one by one back to ground starting from the bottom and toward the top of the array. After reset, even rows are set to all 1's and odd rows are set to all 0's. Then desired values started to be written sequentially from top to bottom. In Figure 2.8, first, a '0' is written into Cell_00 by raising VSS1 and lowering VDDR0. This will also affect the cell in the next row (Cell_10) since it shares the VSS1 and VDDR0 rails, and a '0' is written into Cell_10 as well. However, the disturbed cell is already set to '0' during the reset phase, and hence no erroneous data change occurs. In the next cycle (Figure 2.8, right) , a '1' is

written into Cell_10. Raising VSS2 will also flip the value in the next row, but it was already set to '1' due to the reset process, and again no undesired data change occurs

2.3.2 Write Margin Analysis

Since the write operation is performed through changing VDD and VSS rails of the proposed cell, the write margin must be carefully considered. Basically there are 3 cases of interest. During write we need to lower one of the VDD rails depending on the value to be written. As the VDD rail is shared by same column, this will affect other cells in that column and may flip their values mistakenly. This is called a VDD disturbance and it happens when the VDD rail drops too much. Second, we also raise one of the VSS rails to write a value at the same time, and it will affect all the cells in the same row and may change their values. This is called a VSS disturbance and it happens when VSS rail is increased too much. Finally, if VDD and VSS rails do not change enough, then the write operation itself may fail.

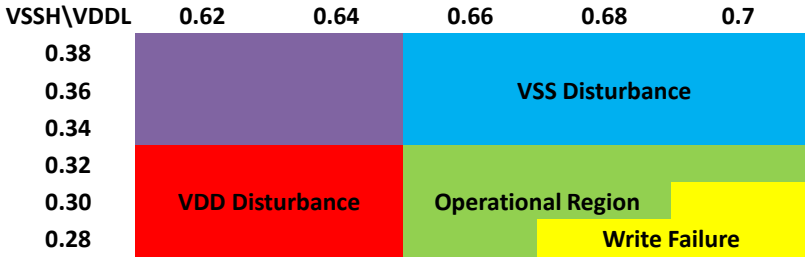


Figure 2.9 Initial write margins of 5T memory design. VDD moves from 1.1V to VDDL while VSS changes from 0 to VSSH.

Figure 2.9 shows the simulated write margins across VDDL and VSSH voltages. Note that higher supply voltage (1.1V) is employed during write operation and the operating voltage is lowered back to normal operating voltage before detection starts. The operational region is defined as the voltage combination that achieves >6 μ/σ for dynamic write margin within 1ns access time, shown in green. Without any assist the operation region is relatively tight and is

only about 40×40mV mainly due to VDD and VSS disturbance. Therefore, the memory needs to be additionally optimized to alleviate the disturbance issues. In simulation, it's observed that weaker transistors are less prone to disturbances and hence different techniques are applied to weaken the pull-up and pull-down transistors (Figure 2.10). First, the opposite VDD rails are lowered for half-selected cells in the same row to maintain the same value in the cell. Also the N-well voltages are simultaneously lowered for the half-selected cells to make the pull-up transistors relatively strong. Extra supply routing for N-well taps does not introduce area overhead but requires one more metal layer. Finally, channel lengths are increased to 50nm instead of using minimum length. The cell area penalty of the increased length is 5.6%, which is already reflected in the 7.2% cell area savings. Figure 2.11 shows the updated write margin plot when the disturbance reduction techniques are applied. VDD and VSS disturbance effects are both reduced now and it provides a significantly enlarged operational region, which is larger than 120mV×180mV. Utilizing multiple supply voltages for read and write assist is a common technique in SRAM designs [28]. If desired, on-chip LDO we can be implemented with minimal overhead (e.g., <0.03mm^2 area and over 96% efficiency in [29]) for each voltage. The voltage switching circuitry increases SRAM macro area by less than 1%. With all these write assist techniques, the overall area saving of the 5T SRAM macro is 5% compared to 6T SRAM.

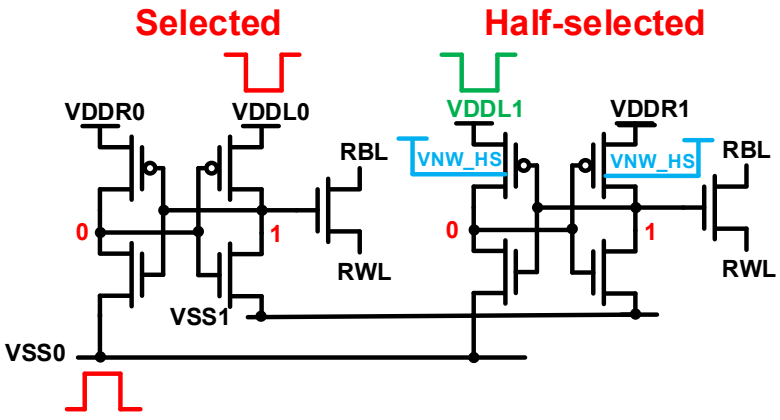


Figure 2.10 Write disturbance reduction techniques.

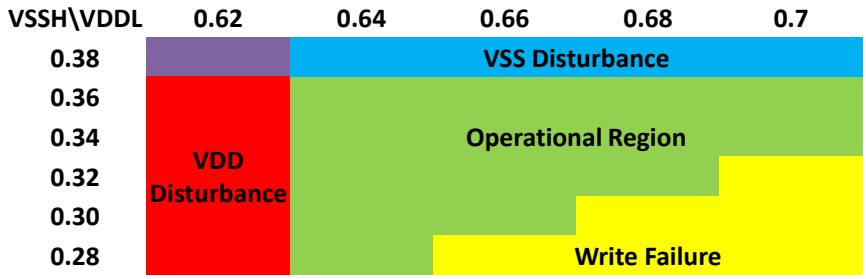
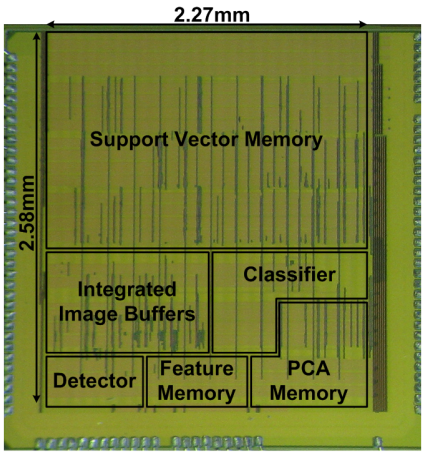


Figure 2.11 Improved write margins of 5T memory with write disturbance reduction techniques at VDD=1.1V.

2.4 Results



Vdd	600 mV
Clock Freq.	100 MHz
Core Area	2.58×2.27 mm^2
Input Video	1280×720 (HD)
Power	23 mW
Throughput	5.5 fps
Efficiency	4.5 nJ/pixel
External Memory Bandwidth	29.53 MB/s

Figure 2.12 Die photo and performance of the face-recognition application.

The 4Mb SRAM is fabricated in TSMC 40nm GP technology for face recognition application. The die photo is shown in Figure 2.12. Figure 2.13(a) shows simulation results regarding the proposed 5T design. Due to the decoupled readout path the 5T SRAM consumes 38% lower read access energy at 0.6V, 100MHz compared to a 6T design, which is a significant amount of saving. The system power consumption is dominated by memory leakage due to leaky

process, and the 38% memory read energy reduction translates to a 3.6% savings in overall energy consumption of the system. Figure 2.13(b) shows the measured minimum read operating voltages for different size. A 4kb array can operate down to 0.39V whereas a 4Mb macro can operate down to 0.52V. The leakage power of the 4Mb 5T SRAM in 40nm technology is 12.1mW, which is 62% of the power consumption of the whole chip. Since 6T SVT SRAM has 5× higher leakage, the overall power saving due to 5T design is 71%. Figure 2.14 shows the diagram of 96kb SRAM unit macro. The unit macro is duplicated to form feature, PCA and support vector memory blocks.

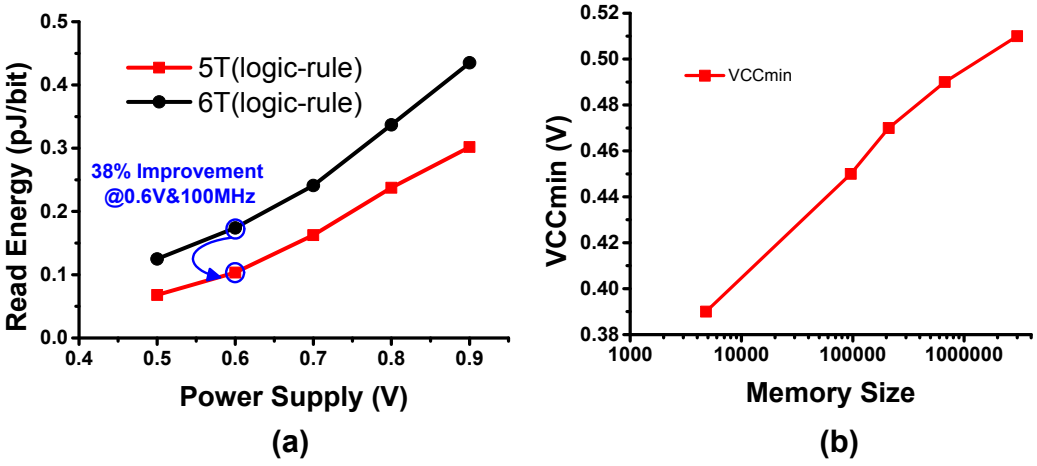


Figure 2.13 (a) Simulated read energy and (b) measured minimum operating voltage.

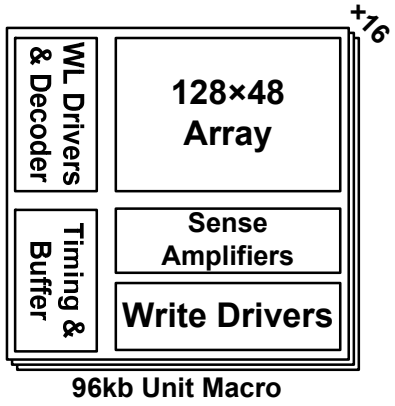


Figure 2.14 Block diagram of 96kb SRAM unit macro with 16 arrays.

Table 2.1 shows comparisons with prior works. From the results one can conclude the proposed 5T has smaller area than conventional 6T while providing much better read margin. Compared with 7T and 8T SRAM [6, 8], the proposed 5T SRAM offers smaller size while maintaining identical read noise margin. The other single-ended 5T [10] has similar area saving, but the shared read/write path undermines the read noise margin. Moreover, the proposed 5T SRAM has the lowest read access energy among the others listed. The write energy is 6.4nJ for a 128×48 array. The energy required to update the entire 5T memory space in the chip is 4.2μJ, whereas the system consumes 4.2mJ to process each image frame. Hence, even if we update memory space before processing each frame, the energy overhead would be less than 1%.

Table 2.1 Comparisons with prior works.

	[8]	[10]	[6]	This Work
Process	65nm	45nm	65nm	40nm
Devices	8T	5T	7T	5T
Voltage	0.35V	0.5V	0.26V	0.6V
Bitcell Size	1.3 x 6T	0.95 x 6T	1.15 x 6T	0.93 x 6T
CLK Frequency	25kHz	250kHz	1.8MHz	100MHz
Read Energy (pJ/bit)	0.88 @ 0.35V	8.8 @ 1V	0.35 @ 0.26V	0.103 @ 0.6V

2.5 Conclusion

A read-optimized 5T memory is proposed to improve voltage scalability and low power consumption. The power-rail-based write scheme and decoupled read paths together significantly improve operations margins as well as provide bit cell area even smaller than other SRAM designs. The 4Mb SRAM is demonstrated fully functional in an energy-efficient face recognition hardware for mobile plat.

CHAPTER 3. 4+2T SRAM for Searching and in-memory-computing Applications

3.1 Introduction

Von Neumann architectures continuously transfer data between memory and computing elements, incurring substantial energy and latency costs that can dominate system power and performance. To minimize this data movement overhead, in-memory-computing allows for data processing inside on-chip memories [30]. In-memory-computing activates multiple rows simultaneously and computes results directly on the bitline. Computation results are immediately available as the memory is accessed, saving clock cycles and interconnect energy. Since memory banks are typically very wide (many bits per word line), it also provides inherently parallel computation.

Conventional 6T SRAM suffers from degraded read noise margin when multiple rows are activated [13], limiting its application to compute-in-memory. 8T SRAM improves read noise margin by decoupling read and write paths but incurs 30% area overhead or more [8]. We propose a 4+2T SRAM cell that uses the N-well as a write wordline, eliminating the access transistors and resulting in a 4T-core memory cell. Two decoupled read paths (2T) significantly improve read noise margin, enabling reliable multi-word activation for logic operations while limiting area overhead to only 12% over 6T SRAM in the same technology. Using dual sense amplifiers, Boolean logic functions (AND, OR, XOR) between the two activated words can be realized. Furthermore, with separated RBL/RBLB and RWL/RWLB, the SRAM can be configured as a BCAM or TCAM, enabling searching operations. The memory cell is designed

using pushed rules in 55nm deeply depleted channel (DDC) technology, which offers a high body coefficient and low process variation.

3.2 4+2T SRAM Cell Design

Figure 3.1 shows the schematic of the proposed 4+2T SRAM cell. The cross-coupled inverters have separated VDD terminals, which serve as WBL and WBLB. Due to the strong body effect in DDC technology, the N-well can be used as WWL. Two decoupled read ports (2T) are used for read and logic operations. In CAM mode, RBL/RBLB and RWL/RWLB are configured as SL/SLB and ML/MLB, respectively. Figure 3.1 also shows the layout and lithographic simulation of the proposed 4+2T cell using pushed rule. Cell area is 265F^2. The WWL (N-well) runs horizontally, and RWL, RWLB, and GND can be shared with adjacent cells. Table 3.1 summarizes the voltages applied on each terminal for basic memory, CAM, and logic operations. Write operation requires two supply voltages, whereas other operations use only one supply voltage.

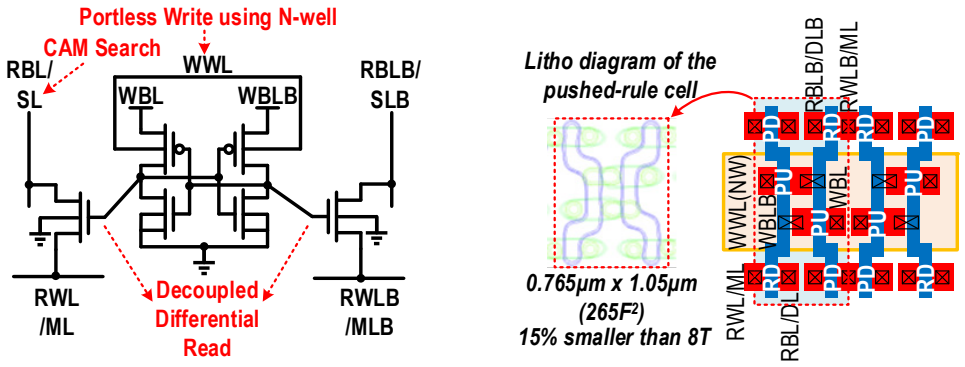


Figure 3.1 4+2T SRAM cell schematic and layout.

Table 3.1 Operation Table.

		WWL	WBL	WBLB	RWL/ML	RWLB/MLB	RBL/SL	RBL/SLB
Memory Operations	WRITE	GND(Sel.) VDDH(Unsel.)	GND(Write0) VDD(Write1)	VDD(Write0) GND(Write1)	VDD	VDD	Floating	Floating
	READ	VDD*	VDD	VDD	GND	GND	Precharge(VDD)	Precharge(VDD)
	HOLD	VDD*	VDD	VDD	VDD	VDD	Floating	Floating
CAM Operations		VDD*	VDD	VDD	Precharge(VDD)	Precharge(VDD)	VDD(Search 0) GND(Search1)	GND(Search 0) VDD(Search1)
Logic Operations	AND	VDD*	VDD	VDD	GND	VDD	Precharge(VDD)	Floating
	OR	VDD*	VDD	VDD	VDD	GND	Floating	Precharge(VDD)
	XOR	VDD*	VDD	VDD	GND	GND	Precharge(VDD)	Precharge(VDD)

*Can also be kept at VDDH.

3.3 Write Operation

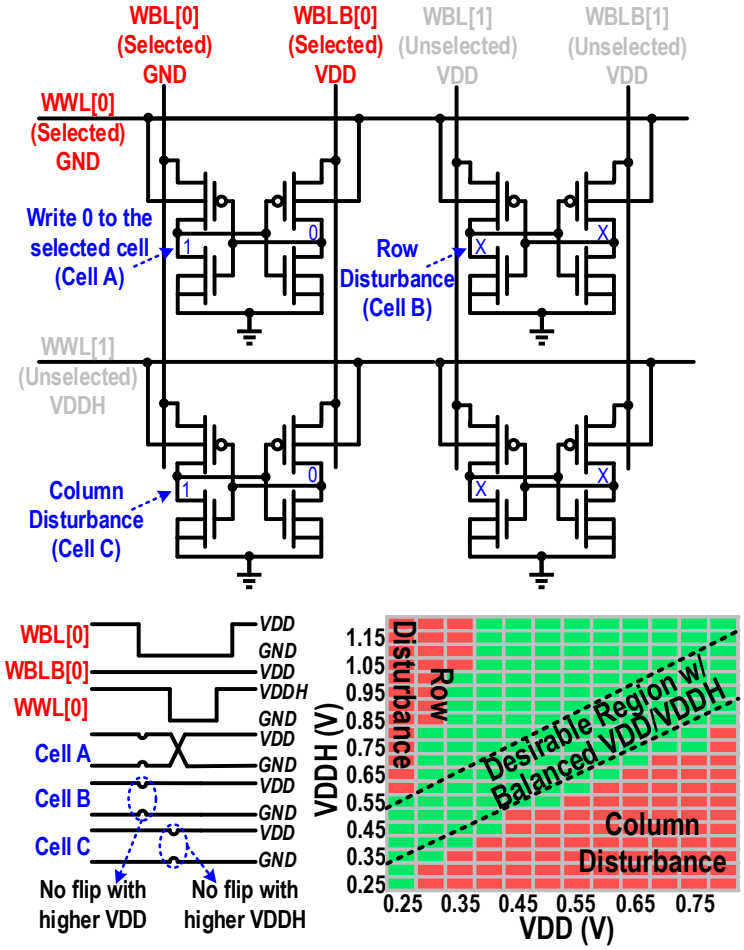


Figure 3.2 Write method and measured write Shmoo plot of 16kb array.

Figure 3.2 shows the write scheme applied to the 4T structure. In standby, both WBL and WBLB are set to VDD, and WWL is at a higher voltage VDDH. To write 0 into a storage node, the selected WBL is lowered from VDD to GND, while WBLB remains at VDD. Once WWL is asserted low, the selected PMOS device becomes much stronger due to its forward body bias. This will short WBL/WBLB with the internal cell node and write into the selected cell. However, there are two types of half-select disturbances that must be taken into account: column-wise and row-wise. Cells on the selected column also have lowered WBL, which can potentially flip the storage nodes. Higher VDDH is thus applied to the WWL of these cells to weaken their PMOS devices and alleviate column disturbances. Conversely, all cells in the selected row have stronger PMOS devices during a write, increasing the chance of un-intended write into their internal nodes. To compensate for this, a higher VDD is applied to their WBL/WBLB to minimize row disturbances. Figure 3.2 also shows the measured margin of cell write, column disturbances, and row disturbances. The green region indicates >5σ combined write margin. Column disturbance occurs at high VDD and low VDDH; row disturbance occurs at low VDD and high VDDH. The operating points centered in the green region (>5σ) have at least ± 200mV VDDH/VDD margin, which is sufficient for robust write.

3.4 Read Operation and Logic-in-memory Operation

Basic read operation is realized with a single decoupled read port similar to a 5T [31] or 7T SRAM [6]. The proposed design uses a differential read port to accelerate read speed and enable logic operations. During a normal read, one pair of RWL/RWLB is activated (pulled low), and one bitline discharges while the other remains high (Figure 3.3, left). The two small column-wise sense amplifiers are connected in parallel to form a larger sense amplifier, accelerating the read operation. For logic operations (Figure 3.3, right), two pairs of wordlines are activated simultaneously. RBL remains high only if both cell nodes (A and B) store 0, and RBL therefore

represents the NAND of A and B. Similarly, RBLB is connected to the complementary nodes and provides the OR of A and B. With the two differential sense amplifiers, NAND/AND and NOR/OR results are simultaneously evaluated. Further, a NOR gate between the two sense amplifier outputs generates the XOR of A and B. All Boolean logic functions are computed in a single read cycle. Since each sense amplifier is small, the area overhead is <5% compared to a normal SRAM, and array efficiency is 65%.

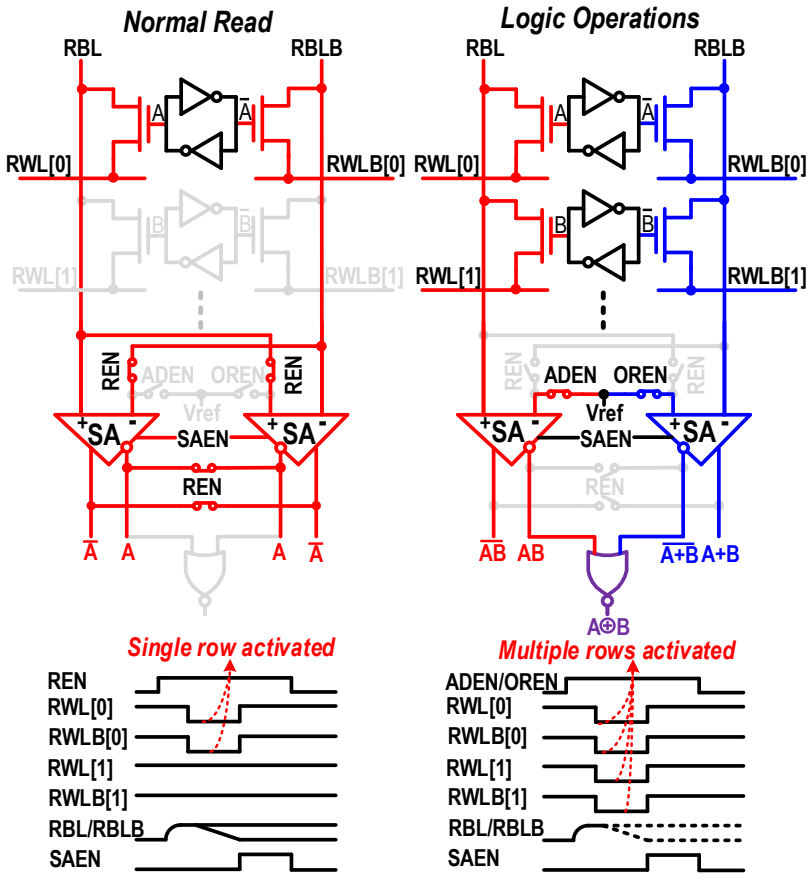


Figure 3.3 Comparison between normal read and Boolean logic operations (AND/OR/XOR).

3.5 BCAM/TCAM Search Operation

Figure 3.4 shows the BCAM/TCAM configurations using the 4+2T SRAM. In CAM mode, the RBL/RBLB supply the search data input SL/SLB, and the RWL/RWLB function as match

lines ML/MLB. For BCAM operation, ML and MLB in a row are shorted together as one matching line. If all the input data match the stored data, ML remains high; otherwise ML discharges. Each ML has a sense amplifier to evaluate the results, similar to a conventional BCAM [32]. Unlike a previous 6T BCAM [13] that requires transposed data storage and two cycles per write, the proposed BCAM stores data in a normal row-wise fashion instead of column-wise. Moreover, read margin is not degraded when multiple rows are activated, in contrast to [13]. A TCAM is realized using dual 4+2T cells. By connecting ML[0] and MLB[1], cell A and cell B can be combined as a single TCAM cell, representing 1/0/X when the AB cells store 00/11/01. The searching and sensing method of TCAM is the same as in BCAM. The BCAM/TCAM uses row-wise differential sense amplifiers with one side connected to a reference voltage.

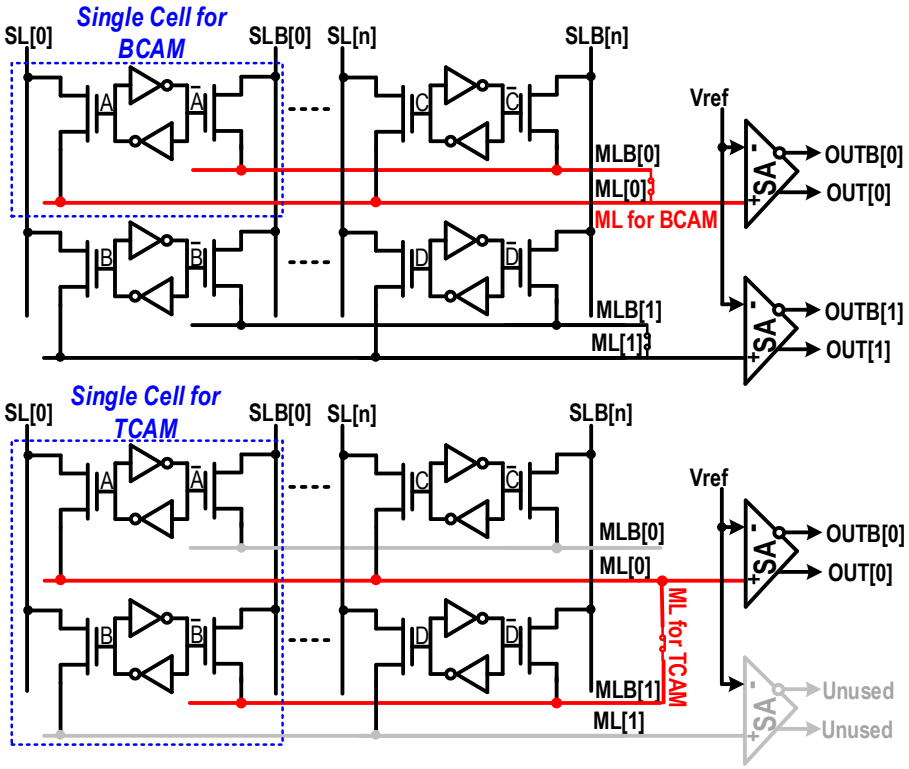


Figure 3.4 BCAM and TCAM configuration.

3.6 Results

The proposed SRAM was fabricated in 55nm DDC technology (die photo in Figure 3.5). The area efficiency is 65% for a 128×128 pushed rule array including all WL/BL/ML peripherals.

Figure 3.6 shows the write frequency and energy across VDD and VDDH. At 0.8V VDD, the write frequency is 600MHz. The minimum supply voltage is 0.25V/0.30V for VDD/VDDH, and the optimal write energy is 4.02fJ/bit at VDD of 0.35V.

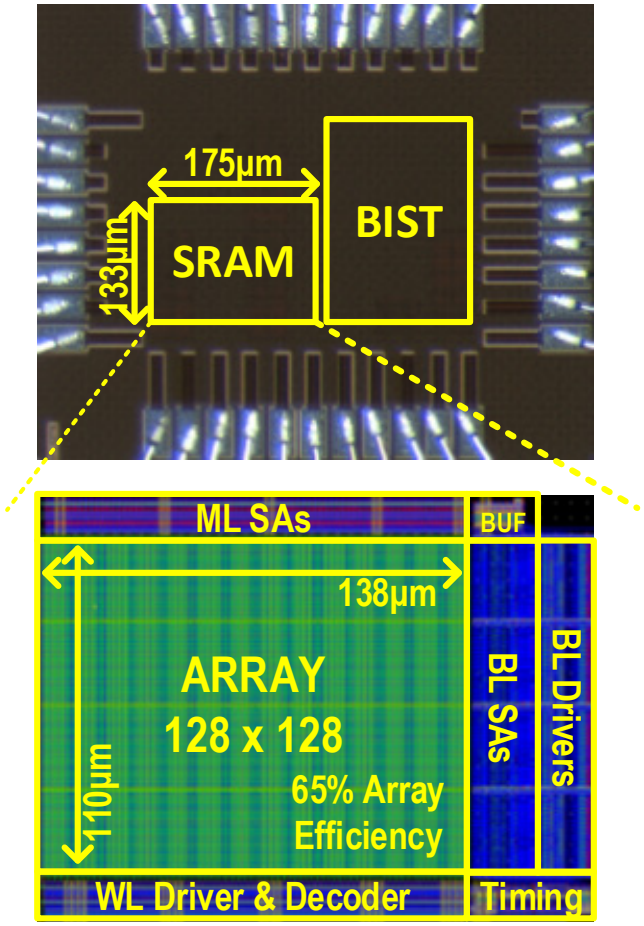


Figure 3.5 Die photo and block diagram.

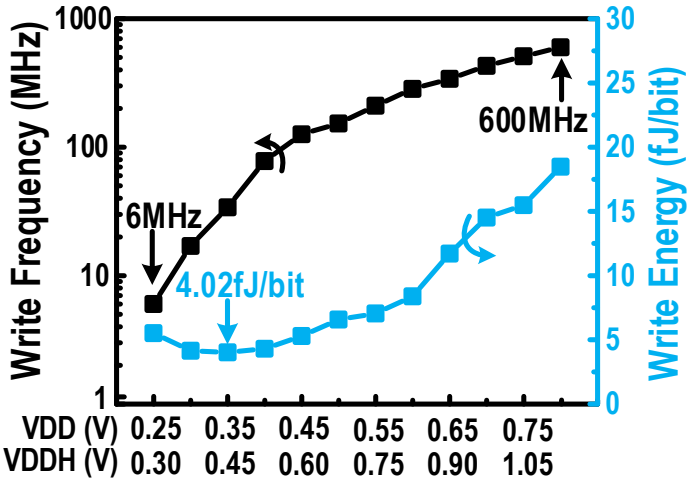


Figure 3.6 Write frequency and energy across VDD/VDDH.

Figure 3.7 shows CAM frequency and energy across VDD. VDDmin is ~0.35V for CAM operation, at which the optimal energy/search is 0.13fJ/bit for BCAM. TCAM has the same frequency as BCAM but 2× the energy/search/bit.

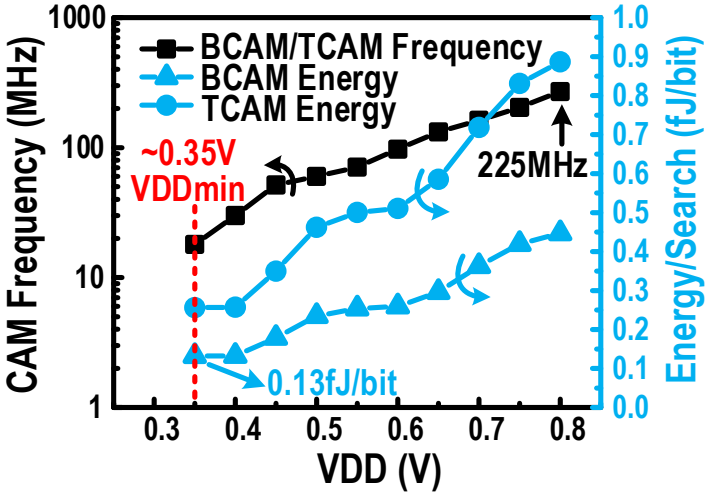


Figure 3.7 BCAM/TCAM frequency and energy across VDD.

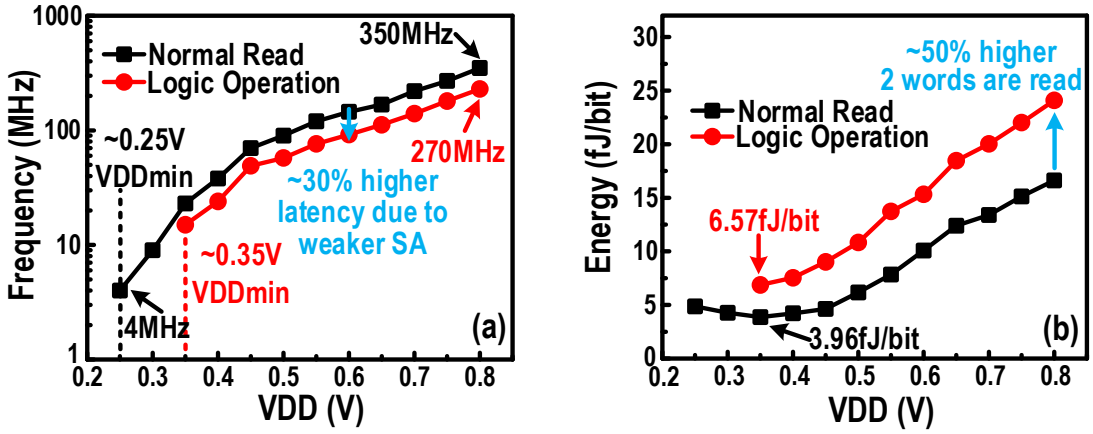


Figure 3.8 Frequency (a) and energy (b) comparison between read and logic operation.

Figure 3.8 shows the optimized frequency and energy across VDD for normal reads and logic operations. VDDmin for read is ~0.25V, whereas it is ~0.35V for logic operations since they employ single-port sensing and half-strength sense amplifiers. The optimal read energy is 3.96fJ/bit at 0.35V; the energy at VDDmin (0.25V) is higher because the leakage energy overhead exceeds the reduction in dynamic energy. The logic frequency is 30% slower and energy/logic operation is 50% higher than that of a normal read operation. However the logic functions operate on 2 words simultaneously instead of a single word as in normal read. Therefore, the total latency (1.3 cycles) achieves 70% savings compared with a conventional 2-cycle read followed by logic. Also, the energy is at least 50% less than that of a 2-cycle read followed by logic.

Figure 3.9 shows the measured VDDmin across temperature for both read/write and hold. Hold VDDmin is ~0.2V at 25°C with 1.6μW leakage power (Figure 3.10). Figure 3.11 shows the within-wafer VDDmin distribution of 20 TT corner dies and the average VDDmin distribution of split wafers in each corner. Table 3.2 compares this work with other decoupled SRAM and CAM works.

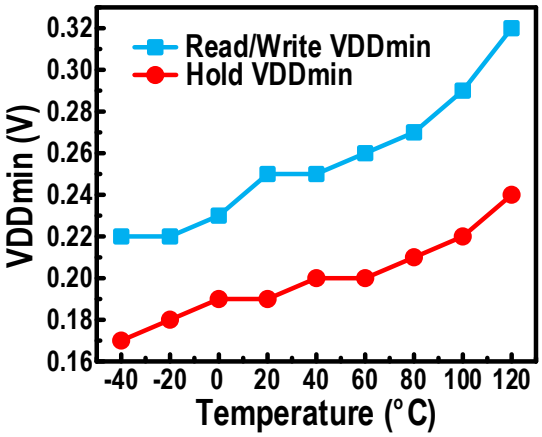


Figure 3.9 VDDmin across temperature.

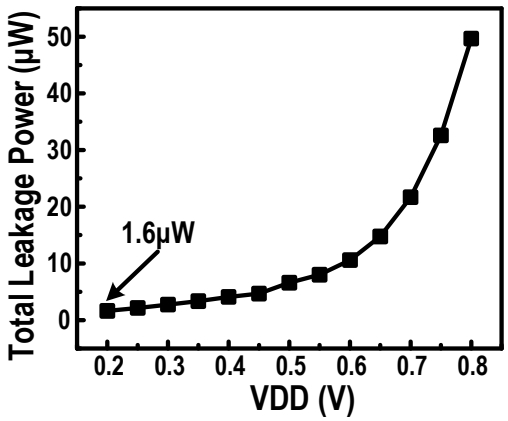


Figure 3.10 Leakage power across VDD.

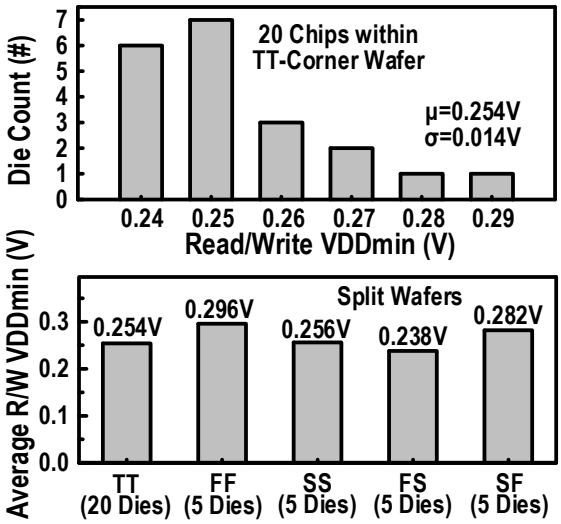


Figure 3.11 Within-wafer and split-wafer VDDmin distribution.

Table. 3.2 Comparison table with other decoupled SRAM and CAM works.

		This work		Decoupled SRAM Work			CAM Work	
				[31]	[6]	[8]	[13]	[32]
Function		SRAM/CAM/Logic		SRAM	SRAM	SRAM	SRAM/CAM/Logic	BCAM
Technology		55nm DDC		40nm	65nm	65nm	28nm FDSOI	32nm
Cell Type		4+2T		5T	7T	8T	6T	11T
Cell Area Scaled to 6T		1.12x		0.93x	1.15x	1.3x	1x	>2x
Pushed-Rule Cell		YES		NO	NO	NO	YES	NO
Array Size		128 x 128 (16kb)		4Mb	256 x 128 (32kb)	256 x 128 x 8 (256kb)	64 x 64 (4kb)	64 x 64 x 4 (16kb)
Array Efficiency		65%		55%	46%	NA	60%	NA
Read/Write VDDmin (V)		0.25		0.38	0.26	0.35		
Write	Freq. (MHz)	600 (0.8V)	6 (0.25V)	NA	NA	0.025 (0.35V)		
	Energy (fJ/bit)[1]	18.5 (0.8V)	5.5 (0.25V)	NA	NA	1240 (0.35V)		
Read	Freq. (MHz)	350 (0.8V)	4 (0.25V)	100 (0.6V)	1.8 (0.26V)	0.025 (0.35V)		
	Energy (fJ/bit)[1]	16.6 (0.8V)	4.9 (0.25V)	103 (0.6V)	44 (0.26V)	880 (0.35V)		
CAM VDDmin (V)		0.35					0.75	0.5
BCAM	Freq. (MHz)	270 (0.8V)	18 (0.35V)				370 (1V)	NA
	Energy (fJ/bit)[2]	0.45 (0.8V)	0.13 (0.35V)				0.6 (1V)	0.3 (0.5V)
Logic	Freq. (MHz)	230 (0.8V)	15 (0.35V)				594 (1V)	NA
	Energy (fJ/bit)[1]	24.1 (0.8V)	6.6 (0.35V)				NA	NA

[1] Divided by word length.
[2] Divided by array size.

3.7 Conclusion

A novel 4+2T SRAM on 55nm deeply depleted channel (DDC) technology is proposed for embedded searching and logic functions. The SRAM cell uses the N-well as the write wordline to perform write operations and eliminate the access transistors, achieving 15% area saving compared to 8T SRAM. The decoupled read paths enables reliable multi-word simultaneous activation to perform Boolean logic functions (AND, OR, XOR). The SRAM can be reconfigured as BCAM/TCAM for searching operations as well and achieves 0.13fJ/search/bit at 0.35V. The chip is fabricated on 55nm DDC technology and achieves 0.25V read/write VDDmin and 0.35V CAM/Logic VDDmin.

CHAPTER 4. Low-power NOR Flash

4.1 Introduction

Increasingly small sensor nodes are ideal for monitoring environmental conditions in emerging applications such as oil exploration. One key requirement for sensor nodes is embedded non-volatile memory for compact and retentive data storage in the event that the sensor power source is exhausted. Non-volatile memory also allows for near-zero standby power modes, which are particularly challenging to achieve at high temperatures when using ultra-low power retentive SRAM due to the exponential rise in leakage with temperature, which rapidly degrades battery life (Figure 4.1). However, traditional NOR flash has mW-level program and erase power [33, 34], which cannot be sustained by mm-scale batteries with internal resistances >10kΩ. To address this issue, an ultra-low power NOR flash is proposed and demonstrated its integration into a complete sensor system that is specifically designed for environmental monitoring under the high temperature conditions experienced when injected in geothermal or oil wells.

The proposed flash design reduces power consumption by using multiple low-power techniques: 1) combined Dickson and ladder pump topology with MIM caps as flying cap; 2) self-adjusting charge pump regulation loop; 3) Ultra-low power voltage and current reference generation circuits. Also, a cross-sampling current sense amplifier is proposed for sensing margin improvement at high temperature. The low power NOR flash is incorporated into a complete mm-scale sensor node system to reduce sleep power and extend battery lifetime.

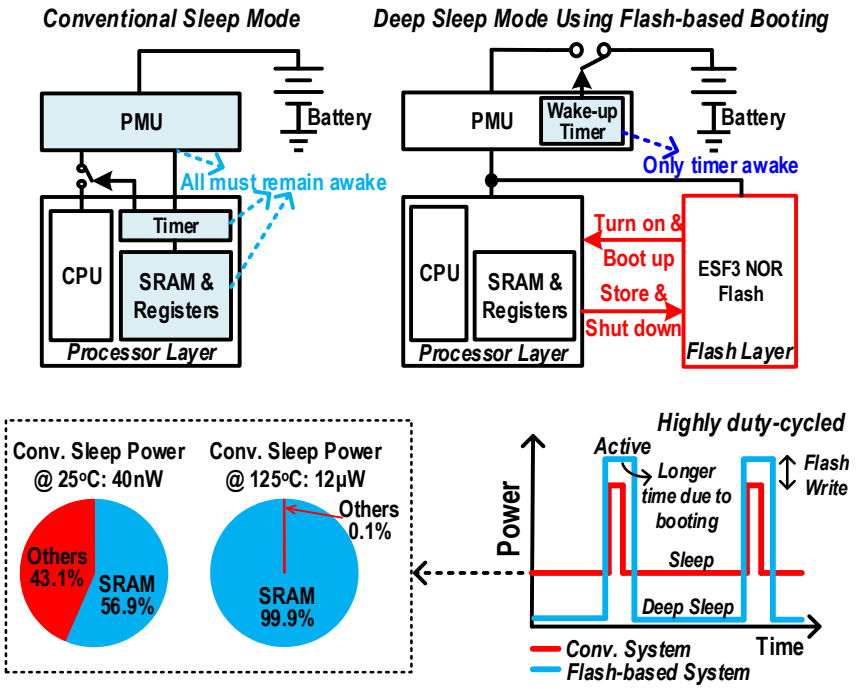


Figure 4.1 SRAM based sensor system keeps PMU, Timer and SRAM awake during sleep; while flash based sensor system requires only a wake-up timer active during sleep.

4.2 High Voltage Generation

4.2.1 Conventional Design

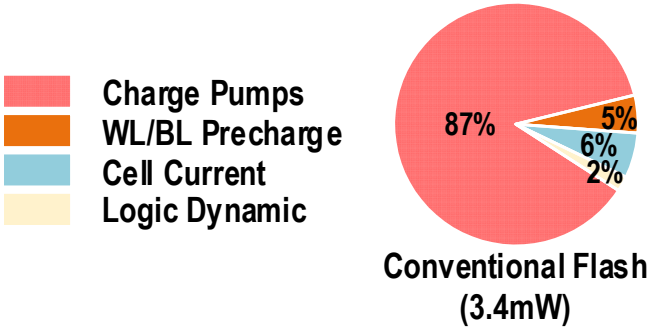


Figure 4.2 Charge pump dominate flash write power.

A charge pump, which generates these high voltages, dominates the write power as shown in Figure 4.2. Dickson pumps have the highest power efficiency. However, the voltage across the flying cap increases with stages in Dickson pump. As a high voltage is applied to the last stage cap, a high-voltage NMOS gate cap must be used as the flying cap for reliability concerns. However, these ultra-thick gate MOS caps have a parasitic/useful cap ratio of 46% (NMOS) and 18% (PMOS) as shown in Figure 4.3. These high ratios result in low pump efficiency, typically less than 30% [34]. MIM capacitors offer very low (1%) parasitic loss, but are limited to <3.6V operation.

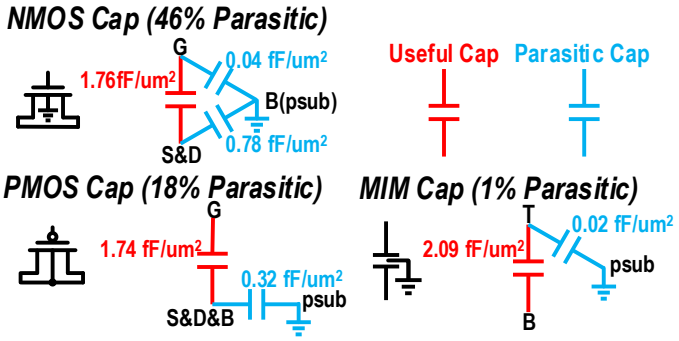


Figure 4.3 Cap parasitic comparison.

4.2.2 Proposed Charge Pumps Design

Embedded split-gate NOR Flash memory requires 4.5V (MV) to provide current for HCI-based program and 13V (HV) for tunneling-based erase operation. The MV pump can still use a MIM cap in a one-stage Dickson pump structure (Figure 4.4) with only Vdd (2.5V) across the flying cap. However, the HV pump cannot use MIM caps in a stacked Dickson pump structure. Therefore, we propose a combined Dickson and Cockcroft-Walton ladder pump (Figure 4.5) to reliably use MIM caps while maintaining high power efficiency. The left part is a Dickson pump structure that doubles the voltage to ~5V in a power-efficient manner. The right part is a ladder pump structure in which the voltage across flying cap remains constant at 2.5V due to stacked

flying caps. Because of the limited voltage across the flying cap, MIM caps can be used. Compared to gate caps, MIM caps have much lower parasitic, which can be less than 1% with no active devices underneath. With a one-stage Dickson pump and a four-stage ladder structure, an output voltage >13V is generated. Finally, a body switch is used to avoid reverse body bias and reduce conduction loss.

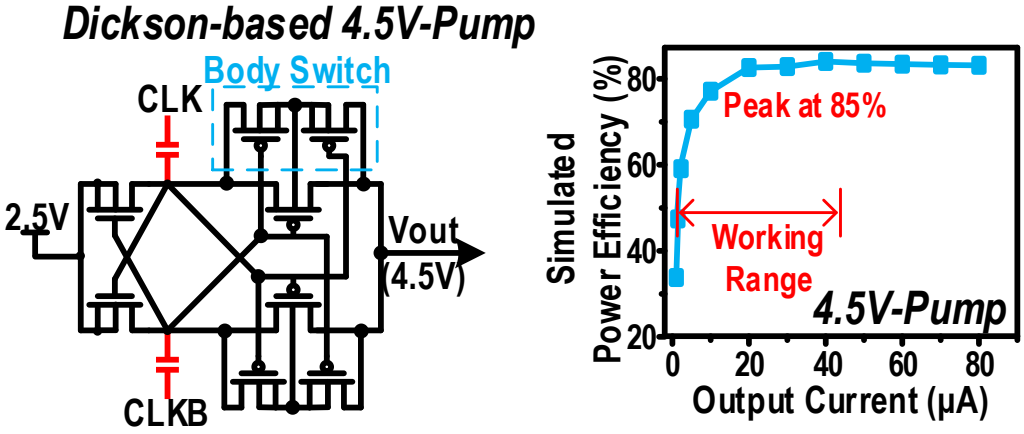


Figure 4.4 MV pump with MIM caps achieving 85% efficiency for the pump loop.

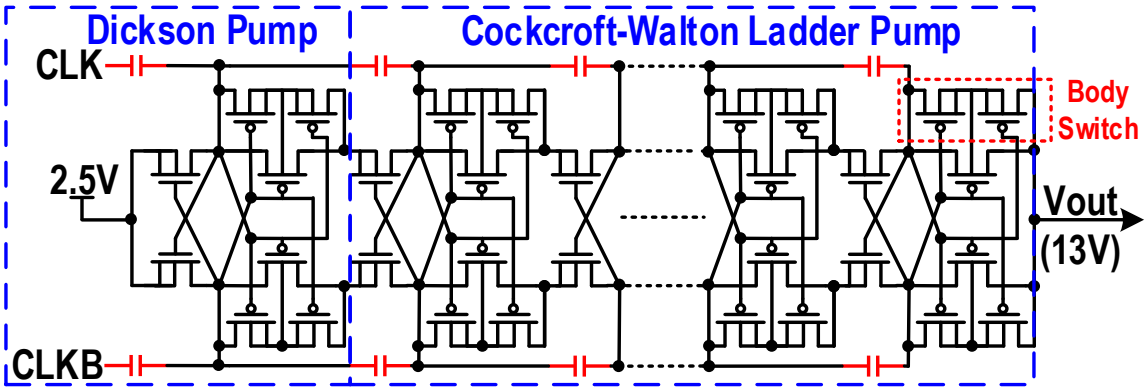


Figure 4.5 Combined Dickson and ladder pump structure.

For the charge pump, we need a regulation loop to stabilize the high output voltage. The regulation circuit uses a dual-Vdd approach (1.2V and 2.5V) to reduce power by 30%. At start-up, erase and program operations require a high VCO frequency (>15MHz) to stabilize Vout,

and hence a high BW amplifier is required. However, read and standby modes do not require a high BW amplifier, and therefore amplifier tail current can be lowered in these modes, reducing the total standby/read power by ~2×. The output voltage is divided down to enable comparison at the lowered voltage domain (1.2V). A high-resistance diode chain divider offers low power but has a long stabilization time. To address this issue, capacitors are placed in parallel with the diode chain to stabilize the loop within 1µs. Figure 4.6 shows that the regulation loop achieves 73% peak power efficiency, which makes ~4× improvement compared with baseline.

Moreover ultra-low-power Vth-based voltage [35] and current reference generation circuits are used to provide constant voltage and current with nW power consumption. These will be discussed in detail later.

4.2.3 Switch-cap Voltage Down Converter

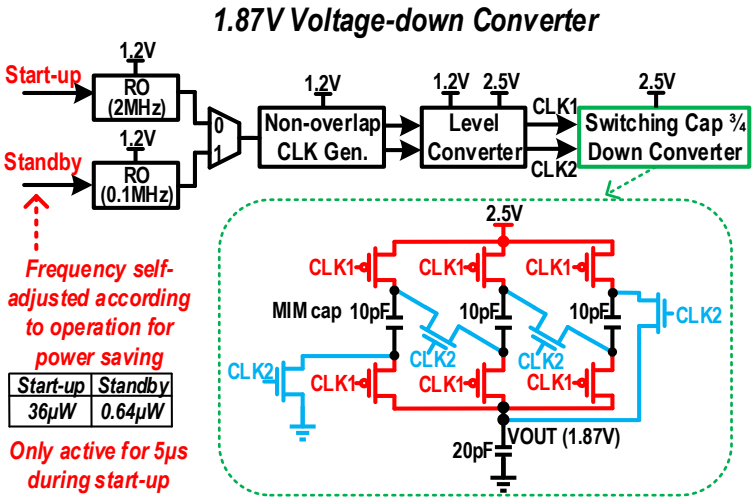


Figure 4.6 Switch-cap based ¾ voltage down converter.

Embedded NOR flash also needs a ~1.8V which is mostly for standby operation. A Switch-cap based ¾ voltage down converter is used to provide the ~1.8V from 2.5V power supply as

shown in Figure 4.6. In the voltage down converter, two ring oscillators are used to provide different clock frequency with optimized power consumption. During start-up, high clock frequency is used to accelerate the stabilization. While during standby, the low-power low-frequency ring oscillator is employed to maintain the output voltage with only 0.64μW standby power.

4.3 High Voltage Delivery

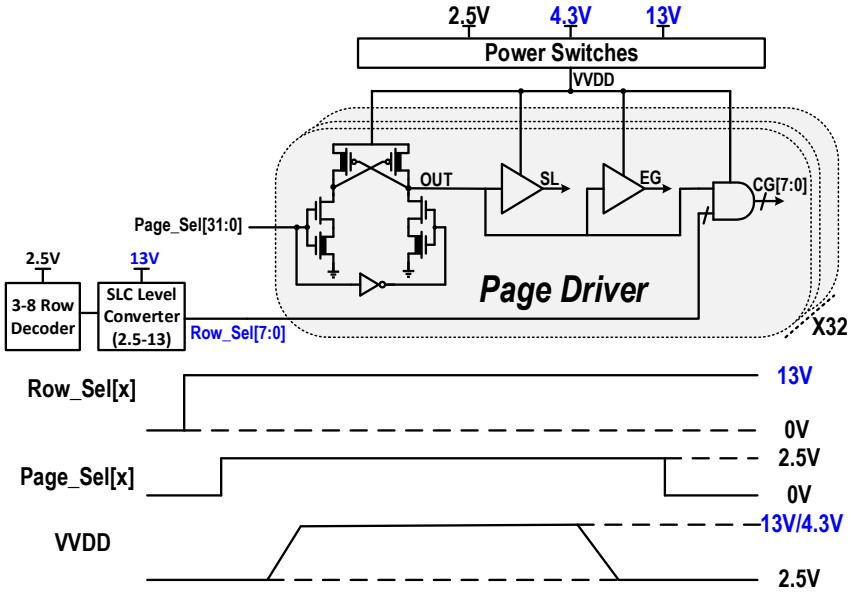
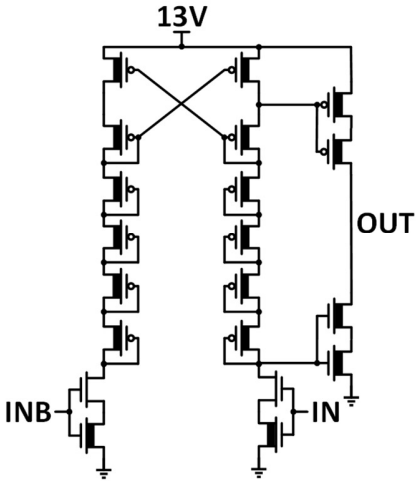


Figure 4.7 Low-power page driver.

Generated high voltages have to be efficiently delivered to the array during write operations. Figure 4.7 shows the page driver circuits. In each on-pitch page drive, there's only one shared DCVSL level converter to save the area. And the power rails of the page drivers are controlled by the shared power switches. The supply of the page driver stays at 2.5V when input changes, and after that the supply switches to high voltage to activate write operation. With no input signal switching at high voltage, dynamic power can be significantly minimized and also the hot switching reliability issue can be avoided.



Mode	2.5 - 13
Power (uW)	3.56
Energy (pJ)	7.11
Leakage (pA)	289

Figure 4.8 Low-power high-voltage level converter.

For those level converters used in shared row decoder and power switch controller, low-power SLC level converters are employed [36], which can reduce the short-circuit-current of the output stage, therefore saving power and avoiding hot-switching.

In power switches, short circuit current between 2.5V and 13V can go to mA range, which has to be avoided. Therefore, non-overlapping pulse generation is used to avoid this short circuit current as shown in Figure 4.9.

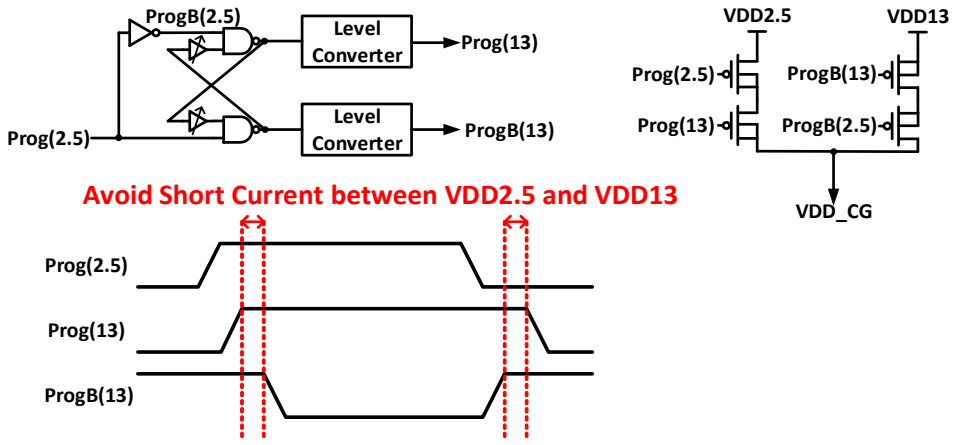


Figure 4.9 Non-overlapping Power Switches.

4.4 Low Power Voltage Reference & Current Reference

4.4.1 Voltage Reference

Voltage reference is required in NOR flash system to provide constant voltage for regulation loop and clamping voltage for current sense amplifiers. The most common way to generate a reference voltage is to use so-called band-gap method [37]. Bandgap circuits take high power because of its innate structure. To achieve low power, one approach is to use a Vth-based voltage reference with devices biased in the sub-threshold region [38, 39]. However, these sub-nW voltage references make use of native transistors, which are potentially at different corners than normal devices due to distinct doping processes, making them more sensitive to process variations. Also, native transistors are not provided by all fabrication technologies [38] and the output reference voltage is too low if an NMOS diode is used. Combining the native NMOS with stacked PMOS diodes can increase the reference voltage [39], but this further enlarges variation across corners.

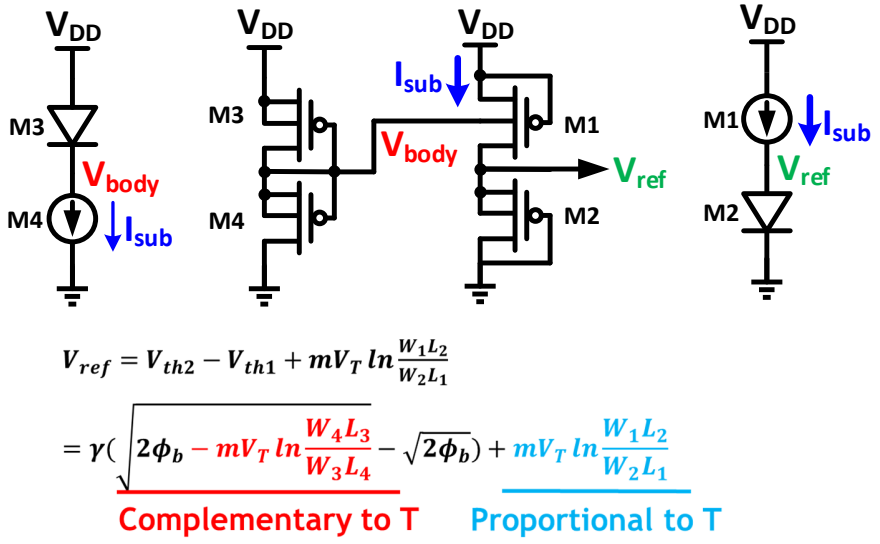


Figure 4.10 Proposed voltage reference circuit and equations.

Figure 4.10 shows a simplified structure of the proposed trim-free voltage reference using only 4 PMOS devices (M1-M4). M1 is forward-biased and provides sub-threshold current flowing through the bottom PMOS diode M2. The current equations of M1 and M2 are expressed as in (4.1). By solving (4.1), Vref can be expressed as (4.3). As M1 and M2 are the same type of PMOS, the difference between Vth1 and Vth2 comes solely from the body bias effect of M1. Random Vth mismatch is kept negligible by upsizing (> 20 μm^2) of all 4 devices in this reference.

$$\boldsymbol{I_R = \ _pC_{ox}\frac{W_1}{L_1}nV_T^2\,exp\left(\frac{0-V_{th1}}{mV_T}\right) = u_pC_{ox}\frac{W_2}{L_2}nV_T^2\,exp\left(\frac{0-V_{ref}-V_{th2}}{mV_T}\right)} \quad \textbf{(4.1)}$$

$$\boldsymbol{I_L = u_pC_{ox}\frac{W_3}{L_3}nV_T^2\,exp\left(\frac{V_{body}-V_{dd}-V_{th3}}{mV_T}\right) = u_pC_{ox}\frac{W_4}{L_4}nV_T^2\,exp\left(\frac{0-V_{th4}}{mV_T}\right)} \quad \textbf{(4.2)}$$

$$\boldsymbol{V_{ref} = V_{th1} - V_{th2} + mV_T\,ln\frac{W_1L_2}{W_2L_1}} \quad \textbf{(4.3)}$$

$$\boldsymbol{= \gamma\left(\sqrt{2\phi_b - mV_T\,ln\frac{W_4L_3}{W_3L_4}} - \sqrt{2\phi_b}\right) + mV_T\,ln\frac{W_1L_2}{W_2L_1}} \quad \textbf{(4.4)}$$

M3 and M4 generate the required body bias for M1. M4 is an off-state PMOS and M3 is a PMOS diode. The current equations of M3 and M4 are expressed in (4.2). As M3 and M4 are also the same type of PMOS, Vth3 and Vth4 are essentially identical. The combination of M3 and M4 provides a body-bias voltage Vbody that tracks Vdd and creates a constant V_{BS} (Vbody-Vdd) for M1, as shown in Figure 4.11. If the current through M3 (I_L) is much larger than the parasitic diode current (Idio) from the source to the N-well of M1, Vref can be expressed by (4.4). The left term of Equation (4.4) is complementary to temperature, whereas the right term is proportional to temperature (Figure 4.10). With proper sizing of the four transistors, the first-order temperature dependency can be cancelled out. Moreover, Vth does not play a role in Equation (4.4) because each pair (M1/M2 and M3/M4) uses the same type of PMOS, thus significantly reducing process variation. Since Idio is not well modeled, I_L is designed to be 3 orders of magnitude larger than Idio to minimize the effect of Idio. Proper sizing of these transistors can be determined using a global optimization tool.

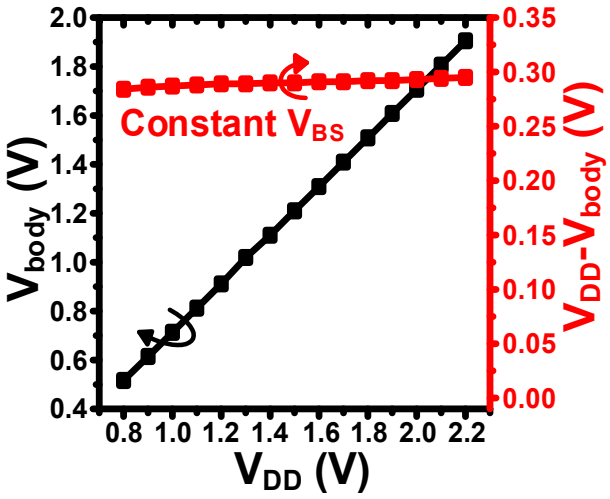


Figure 4.11 V_{body} tracks V_{dd} change and creates constant V_{BS} for M1.

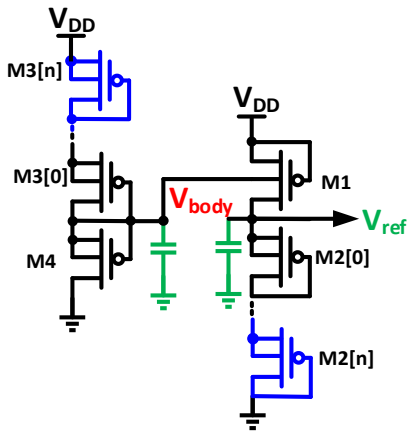


Figure 4.12 Proposed voltage reference generator with stacked PMOS diodes.

As shown in Figure 4.12, stacked PMOS diodes can replace M2 and M3 to generate a higher reference voltage, and multiple voltage levels can be generated in this manner. Three stages of PMOS diodes are used in our design to realize an approximately 1V output reference voltage. MIM capacitors C0 and C1 (both set to 1.78pF) are used to isolate the reference voltage from high-frequency power supply noise.

Figure 4.13 compares simulated reference voltage distributions across corners for the proposed design as well as designs from [38] and [39]. The proposed design achieves < 4% inaccuracy across all corners, whereas [38] and [39] vary up to 10% and 19%, respectively.

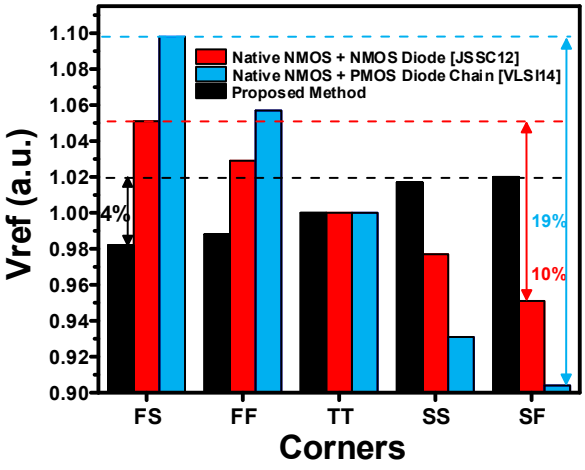


Figure 4.13 Comparison of Vref simulation at all corners among the proposed design, [38], and [39].

We first verified the proposed voltage reference in 180nm technology before incorporated into 90nm Flash technology. Figure 4.14 shows the die photo; the proposed voltage reference occupies an area of 4880μm^2 (80μm x 61μm) with this area dominated by the two MIM capacitors, C0 and C1. Sixty chips from 3 different wafers in 180 nm CMOS were tested. One wafer was in a typical corner with thin top-metal, another was found to be at a slow corner with ultra-thick top-metal, and the third was at a fast corner with ultra-thick top-metal. All measurements are reported without trimming.

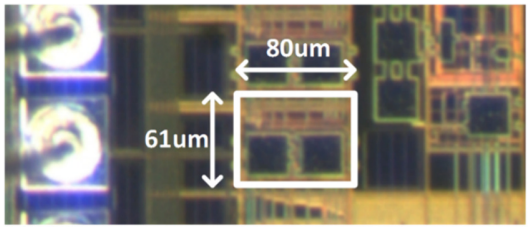


Figure 4.14 Die Photo in 180 nm CMOS.

Figure 4.15 shows the measured reference voltage across temperature for all 60 chips. From −40°C to 85°C, the temperature coefficient of the typical wafer ranges from 48ppm/°C to 104ppm/°C, and those of the fast and slow wafers are 55.2−124ppm/°C and 56.1−117ppm/°C, respectively. The reference voltage distributions at 25°C of the 3 different wafers are shown in Figure 4.16. Without trimming, the typical wafer shows a mean value of 986.2mV and standard deviation of 2.6mV. The average voltage difference between the fast and slow wafers is 3.6% (1.9% σ/μ), matching simulation and providing sufficient accuracy for many key circuit applications.

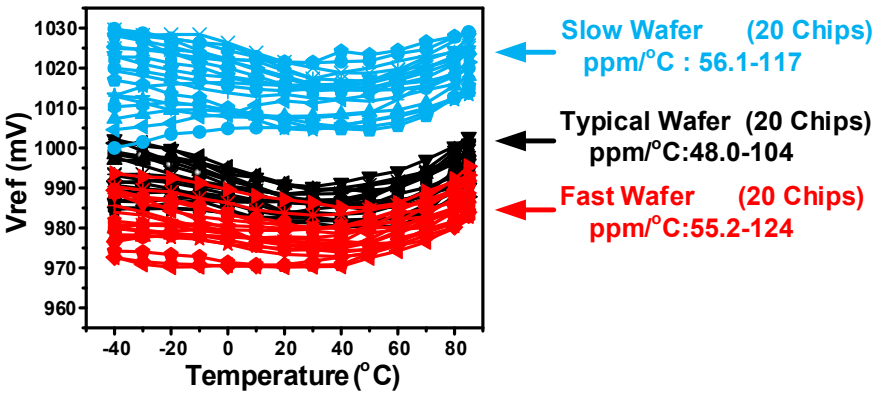


Figure 4.15 Measured Vref across temperature for 3 wafers in 3 different corners.

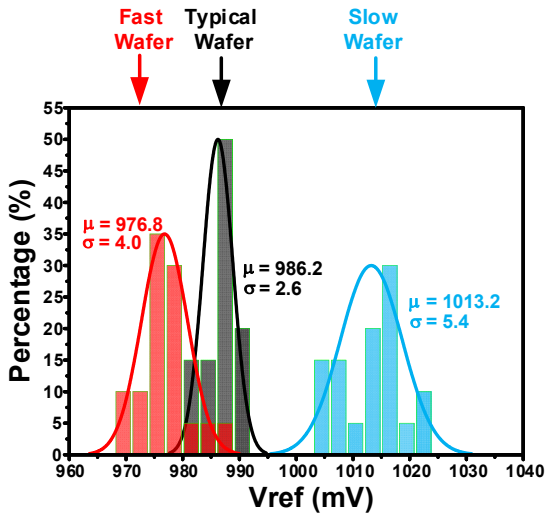


Figure 4.16 Distribution of Vref on 3 different wafers.

Figure 4.17 shows the measured sensitivity of reference voltage to power supply voltage. Line sensitivity is 0.38%/V from 1.2V to 2.2V. Figure 4.17 also shows the measured power supply rejection ratio (PSRR) from 10Hz to 10MHz. High-frequency PSRR is −56dB, which can be further improved with larger loading caps C0 and C1.

Figure 4.18 shows the measured power consumption across supply voltage and temperature. The output reference voltage is approximately 1V with 3 stages of stacked PMOS diodes. The power supply can be reduced to 1.2V while maintaining this approximately 1V reference voltage. To lower the minimum power supply, fewer stages of PMOS diodes can be used, but the output reference voltage will be lowered as well. At 25°C and 1.2V, the power consumption is 114pW, which is suitable for low-power sensor and IoT applications.

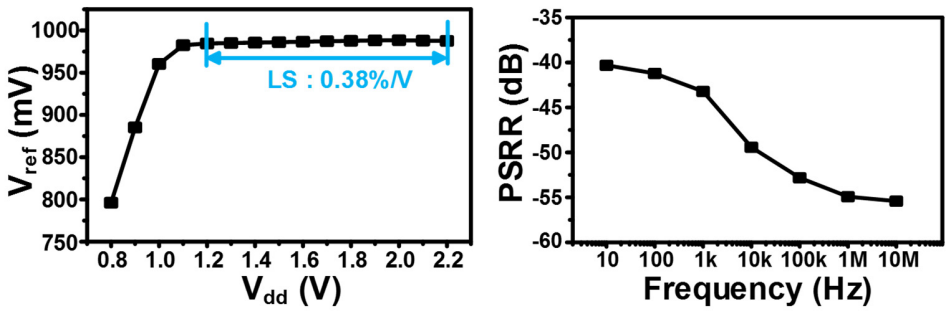


Figure 4.17 Measured line sensitivity and PSRR.

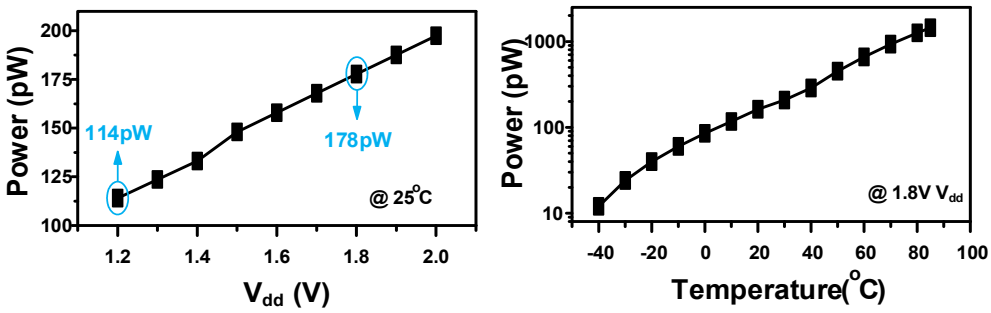


Figure 4.18 Measured power across Vdd and temperature.

Figure 4.19 shows the combined uncertainties of process, voltage and temperature together. Three sigma untrimmed within-wafer process variation is around 0.78% inaccuracy. 100 degree temperature change will accumulate 0.75% inaccuracy. And 1V Vdd change will have 0.38% inaccuracy. The total uncertainty with PVT is around 1.9% for this work. Table 4.1 summarizes the results of the proposed sub-nW trim-free voltage reference and compares them with previous works.

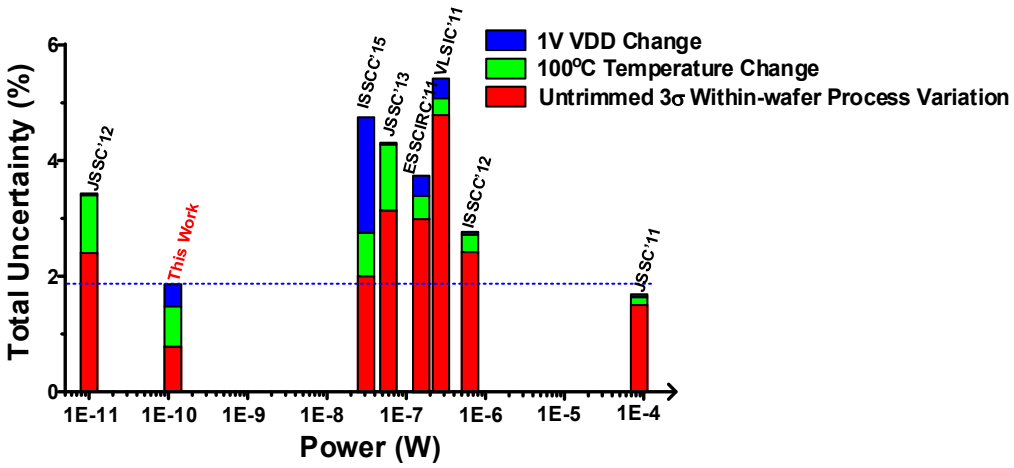


Figure 4.19 Comparison of combined uncertainties with other works.

Table. 4.1 Comparison table of voltage reference.

Parameters	This Work	[38]	[40]	[41]	[42]	[37]
Process (nm)	180	180	130	180	130	160
Power (nw)	0.114	0.006	32	52.5	170	99000
Min. V_{dd} (V)	1.2	0.5	0.5	0.7	0.75	1.62
V_{ref} (V)	0.9862	0.3268	0.498	0.548	0.256	1.0875
Within-Wafer Untrimmed σ/μ (%)	0.26	0.8	0.67	1.05	1	0.5
Wafer-to-Wafer Untrimmed σ/μ (%)	1.9 (3 wafers)	NA	NA	NA	NA	0.3* (2 wafers)
Temp. Range (°C)	-40 ~ 85	-20 ~ 80	0 ~ 80	-40 ~ 120	-20 ~ 85	-40 ~ 125
TC (ppm/°C)	48.0 ~ 124	54.1 ~ 176.4	75	114	40	5 ~ 12
LS (%/V)	0.38	0.044	2	NA	0.35	NA
PSRR (dB)	-42/-56 (100Hz/10MHz)	-49/-55 (100Hz/10MHz)	-40	-56 (100Hz)	-93	-76 (DC)
Area (um^2)	4880	1425	26400	24600	70000	120000
Type	Vth	Vth	Bandgap	Bandgap	Bandgap	Bandgap
Chips Measured	60 chips in 3 wafers	14 chips	6 chips	9 chips	NA	61 chips in 2 wafers

4.4.2 Current Reference

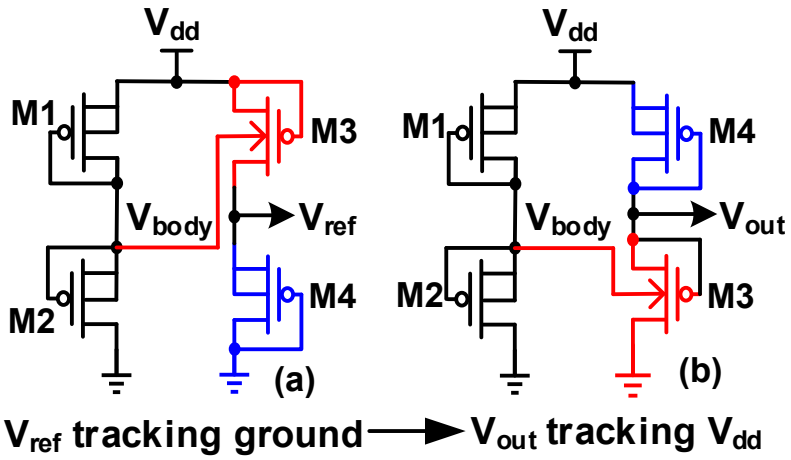


Figure 4.20 Voltage generation circuits for ground tracking (a) or Vdd tracking (b).

Current references are also fundamental elements in low-power NOR flash system, which demands low-power nano-ampere range current references that operate across a wide temperate range to satisfy reliable operation under a range of environmental conditions. Conventional beta-multiplier current references require very large polysilicon resistors or resistor-like MOSFET [43] to generate a nA-reference current, incurring substantial area overhead. Sub-threshold current reference [44] using both NMOS and PMOS suffers from process variation and limited temperature range due to inaccurate leakage models at high temperature, particularly for Nwell leakage. This leads to high post-fabrication calibration costs to tighten the distribution of the generated reference current.

The proposed current reference is based on a Vth-based voltage reference design that we discussed above. In this voltage reference, shown in Figure 4.20(a), an off-state M2 and diode M1 provide the required body bias for M3 to generate supply-insensitive sub-threshold current flowing through the bottom PMOS diode M4. With proper sizing of the four PMOS transistors, the first-order temperature dependency of the generated reference voltage Vref can be cancelled

out. Vref is insensitive to Vdd while tracking the ground voltage. By swapping the positions of M3 and M4 (Figure 20(b)), the generated Vout can instead track Vdd, maintaining a constant Vdd-Vout. As shown in Figure 21, we use this Vout to control the gate voltage of a PMOS M5. As shown in Figure 22, both Vout and Vbody track Vdd and therefore Vdd-Vout is kept constant for Vdd >1.4V. Since Vdd-Vout is constant, M5 |Vgs5| is supply insensitive (Figure 22) and Iref (M5 drain current) has very weak sensitivity to Vdd.

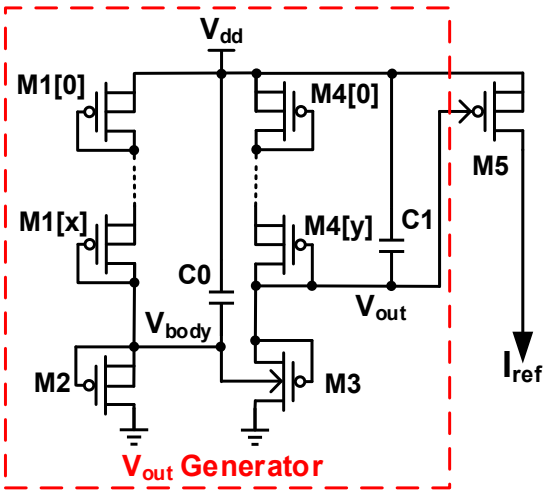


Figure 4.21 Schematic of the proposed design.

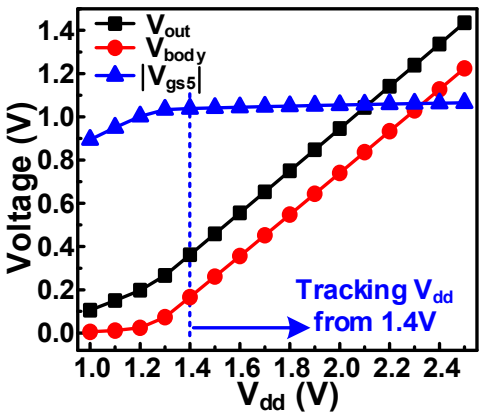


Figure 4.22 Simulated V_{out}, V_{body} and $|V_{gs5}|$ tracking V_{dd} change.

Proper sizing of M1/M2/M3/M4 is necessary to compensate for the temperature coefficient of Iref. The current equations of M1/M2 and M3/M4 are expressed as (4.5) and (4.6), respectively, below:

$$\boldsymbol{I_L = \ _pC_{ox}\frac{W_1}{L_1}nV_T^2\,exp\left(\frac{(V_{dd}-V_{body})/x-V_{th1}}{mV_T}\right) = u_pC_{ox}\frac{W_2}{L_2}nV_T^2\,exp\left(\frac{0-V_{th2}}{mV_T}\right)} \quad \textbf{(4.5)}$$

$$\boldsymbol{I_R = u_pC_{ox}\frac{W_4}{L_4}nV_T^2\,exp\left(\frac{(V_{dd}-V_{out})/y-V_{th4}}{mV_T}\right) = u_pC_{ox}\frac{W_3}{L_3}nV_T^2\,exp\left(\frac{0-V_{th3}}{mV_T}\right)} \quad \textbf{(4.6)}$$

By solving (4.5) and (4.6), Vgs5 can be expressed as (4.7):

$$\boldsymbol{|V_{gs5}| = y\gamma\left(\sqrt{2\phi_b - |V_{gs5}| + xmV_T\,ln\frac{W_2L_1}{W_1L_2}} - \sqrt{2\phi_b}\right) + ymV_T\,ln\frac{W_3L_4}{W_4L_3}} \quad \textbf{(4.7)}$$

in which *x* and *y* are the stage number of M1 and M4, respectively. By properly selecting values for *x* and *y*, |Vgs5| can be biased close to the zero temperature coefficient (ZTC) voltage of M5's saturation drain current. This minimizes the temperature coefficient of Iref in first order.

$$\boldsymbol{I_{ref} = u_pC_{ox}\frac{W_5}{L_5}(|V_{gs5}| - V_{th5})^2} \quad \textbf{(4.8)}$$

Equation (4.8) describes Iref in the saturation region. Because PMOS mobility μ_p is complementary to temperature and determines the temperature dependence of drain current in saturation region, |Vgs5| must be slightly proportional to temperature for second order compensation. In equation (4.7) if $W_3L_4 > W_4L_3$, the temperature dependence of the right term is positive; otherwise it is negative. A similar trend holds for the left term containing W_2L_1/W_1L_2 in the square root. Therefore, sizing M1/M2/M3/M4 allows tuning of the temperature dependence of |Vgs5| to a proportional value (PTAT) that minimizes the overall temperature coefficient of Iref. The proposed design has 4 PMOS widths and lengths for a total of 8 parameters which need to be determined. A global optimization flow is used to find the optimum sizes to compensate temperature sensitivity. Figure 4.23 shows the simulated waveform of Vout, Vbody, and Vgs5 with temperature ranging from -40˚C to 120˚C. |Vgs5| is close to the ZTC voltage with a slightly positive temperature correlation. Process variation is suppressed due to the exclusive use of

PMOS transistors, which are also upsized. Two 1.78pF MIM capacitors C0 and C1 (1.78pF) weaken the impact of high-frequency power supply noise as shown in Figure 4.21.

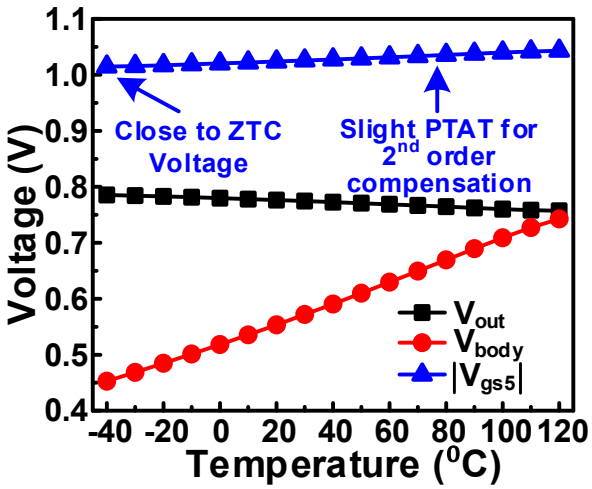


Figure 4.23 Simulated V_{out}, V_{body} and $|V_{gs5}|$ across temperature.

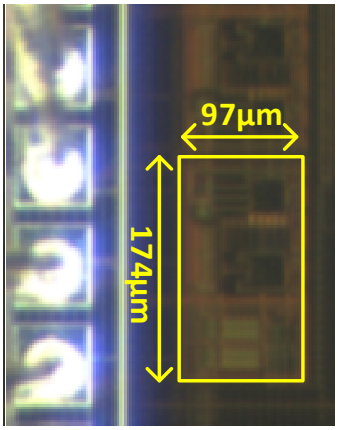


Figure 4.24 Die Photo in 180nm technology.

The proposed voltage reference is also verified in 180nm technology. Figure 4.24 shows the die photo. The area of the proposed circuit is 16878µm^2 (97µm x 174µm). Measurements include 32 chips from 5 wafers, including one at TT corner (16 dies) and four corner wafers (4 dies for each of FF, SS, SF, and FS). Fig. 4.25 shows measured Iref across temperature without trimming for all 32 chips.

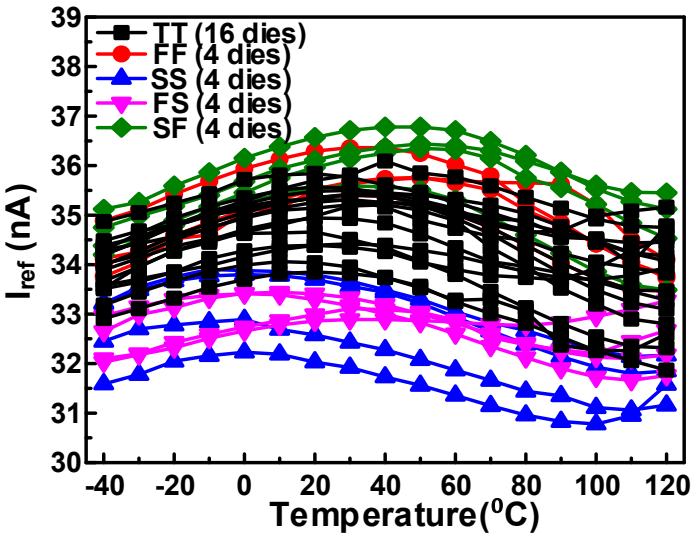


Figure 4.25 Measured Iref across temperature for 5 wafers in 5 different corners.

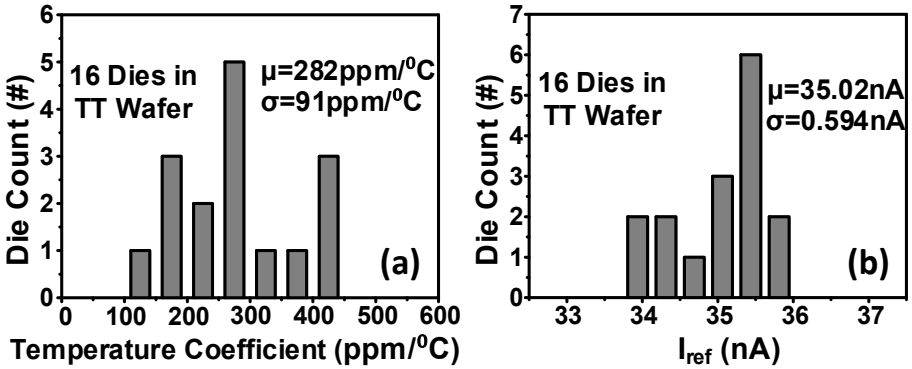


Figure 4.26 Measured temperature coefficient distribution of 16 dies in TT wafer (a). Measured Iref distribution of 16 dies in TT wafer at room temperature (b).

Figure 4.26(a) illustrates the temperature coefficient distribution of the 16 chips at TT corner. The average temperature coefficient is 282ppm/˚C without trimming. The average Iref at room temperature for 16 chips is 35.02nA with 0.594nA standard deviation (Figure 4.26(b)), which is 1.6% σ/μ for within-wafer variation. Figure 4.27 shows the measured average Iref at each temperature point for different corners. With higher PMOS threshold voltage (TT, SS, FS), second order temperature compensation is observed at ~100˚C, extending the working

temperature range. The average temperature coefficient at TT, SS, and FS corners are all below 300ppm/˚C; while FF and SF corners have slightly worse untrimmed temperature coefficient (Figure 4.28 left). The average Iref at room temperature has ±4.7% difference across the corner wafers (Figure 4.28 right).

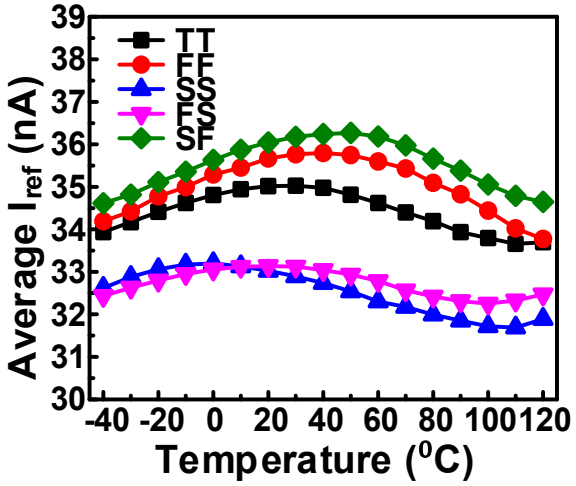


Figure 4.27 Measured average Iref across temperature for 5 corner wafers.

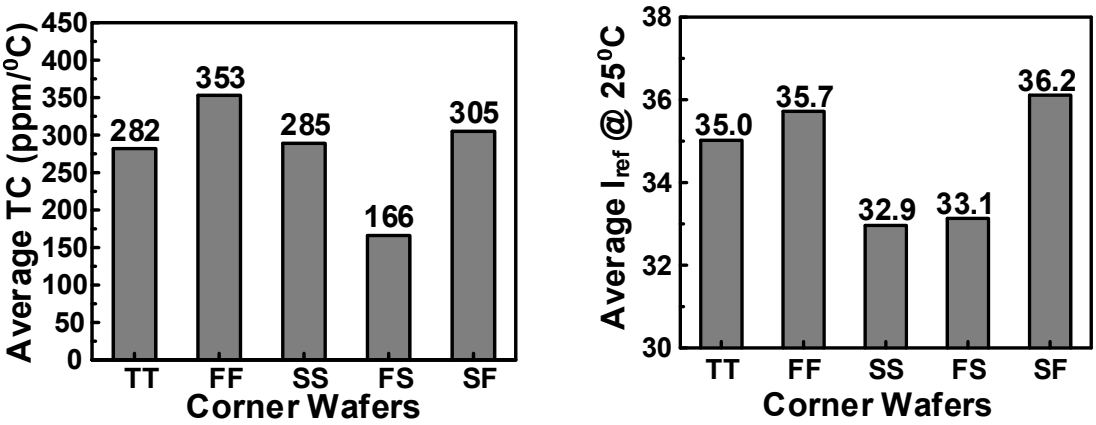


Figure 4.28 Measured average temperature coefficient in each corner wafer and measured average Iref comparison among 5 corner wafers.

Measured line sensitivity is approximately 3% for Vdd above 1.5V (Figure 4.29). Figure 4.30 left shows the measured power across Vdd at room temperature. At 1.5V, the proposed

circuit consumes 1.02nW in TT corner at room temperature. In the worst corner (FF), power remains below 1.7nW at 1.5V. Furthermore, even at 120˚C, power consumption is below 25nW across all corners (Figure 4.30 right), which is sufficiently low for many applications.

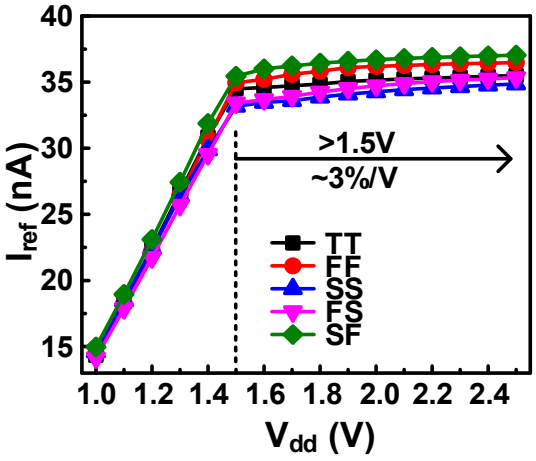


Figure 4.29 Measured line sensitivity for 5 corner wafers.

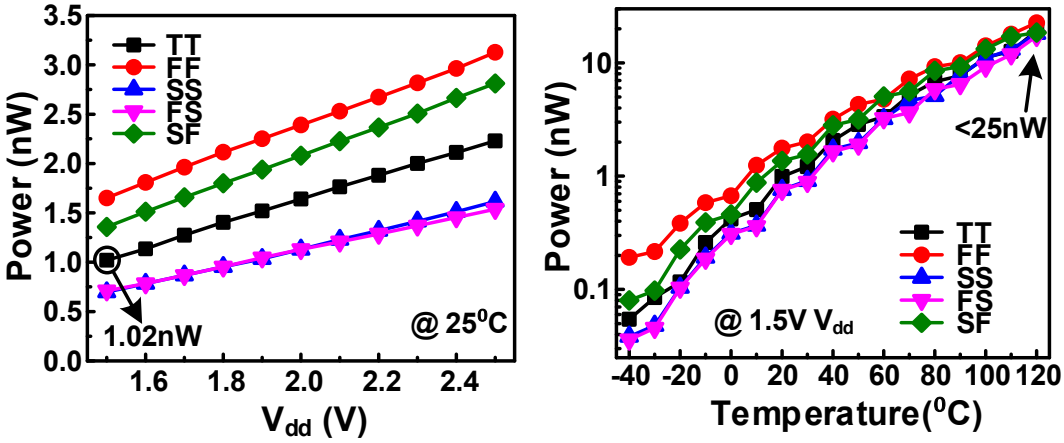


Figure 4.30 Measured power across Vdd and temperature for 5 corner wafers.

Figure 4.31 shows the accumulated uncertainties of PVT variation for state-of-the-art current references. For the proposed reference the 3σ untrimmed within-wafer process variation is ~4.8%, a 100˚C temperature change induces a 2.8% variation; and 0.4V supply voltage change

incurs 1.2% deviation. Thus the total PVT-induced uncertainty is ~8.8% for this work, which is the smallest among relevant works. Table 4.2 summarizes measured results of the proposed current reference and compares them with previous works.

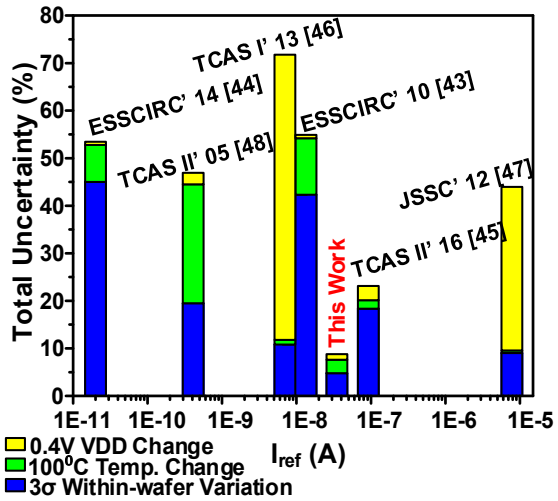


Figure 4.31 Accumulated uncertainty comparison with other works.

Table. 4.2 Comparison table of current reference.

	This Work	[43]	[45]	[46]	[47]
Technology (nm)	180	350	180	350	180
V_{dd} (V)	> 1.5	> 1.3	> 1.25	5	> 1
I_{ref} (nA)	35	9.95	92.3	9	7810
Power (nW)	1.02	88.5	670	4171	1400
Temperature Range (°C)	-40 ~ 120	-20 ~ 80	-40 ~ 85	0 ~ 80	0 ~ 100
TC (ppm/°C)	282 (146 ~ 428)	1190	177	57	24.9
Line Sensitivity (%/V)	3	0.046	7.5	150	86 [1]
Within-wafer Variation (σ/μ)	1.6%	14.10%	6.12%	3.60%	3% [2]
Wafer-to-wafer Difference	± 4.7% [3]	NA	NA	NA	NA
Trimming	No	No	NA	NA	No
Chip Area (mm^2)	0.017	0.12	0.0013	0.0081	0.023
Samples	32 from 5 wafers in split corners	15 from 1 wafer	10 from 1 wafer	3 from 1 wafer	10 from 1 wafer

[1] Value w/o BGR, extracted from Fig. 10(a)
[2] σ is estimated according to ± 4.5% I_{ref} difference
[3] Difference among corner wafers

4.5 Array Organization

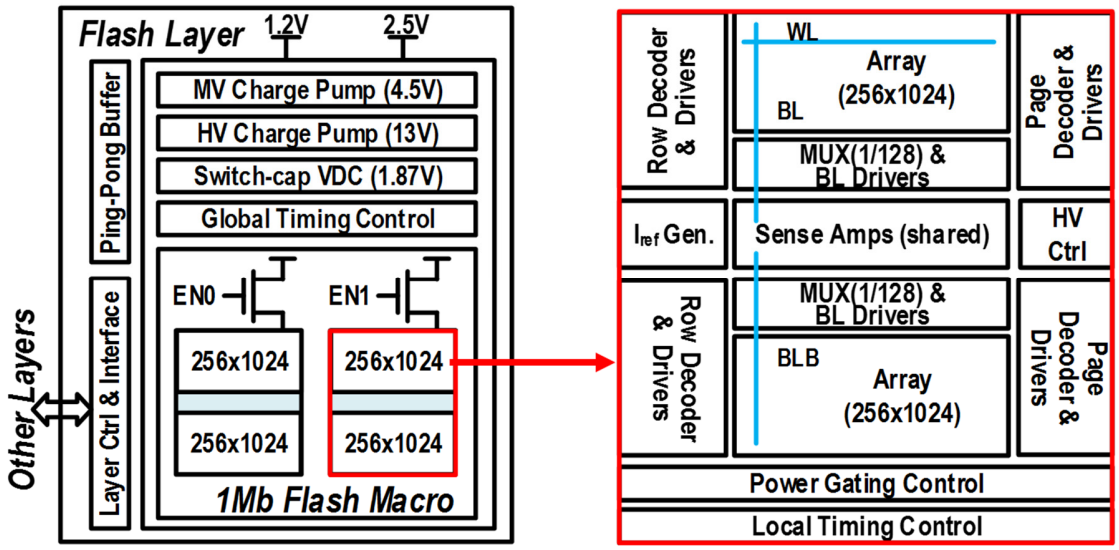


Figure 4.32 Block diagram of the 1Mb flash macro and array organization.

Figure 4.32 shows the block diagram of the 1Mb flash macro, which is separated into two banks, each with its own power gating control. When one bank is active, all peripherals in the other bank are power-gated. Each bank has two arrays with 256*1024 cells. The current sense amplifiers, reference current generation, and high voltage switches are shared by the two arrays. Page-wise erase mode requires 8Kb; only 8 bits are programmed or read at one time to lower instantaneous power. With the low-power charge pump and optimized array organization, the inherent cell current consumes >50% of the total power, reducing power overhead from 1560% to 81%.

4.6 Sense Amplifier Design

4.6.1 Margin Degradation at High Temperature

Figure 4.33 shows the simulated read cell current across temperature. The read current of a programmed cell increases with temperature due to Vth decreasing; while that of an erased cell decreases because of degraded mobility. As a result, the read current ratio between the two states thus reduces from 8× to 5× as the temperature increases from 25°C to 125°C, complicating sense amplifier design at high temperatures.

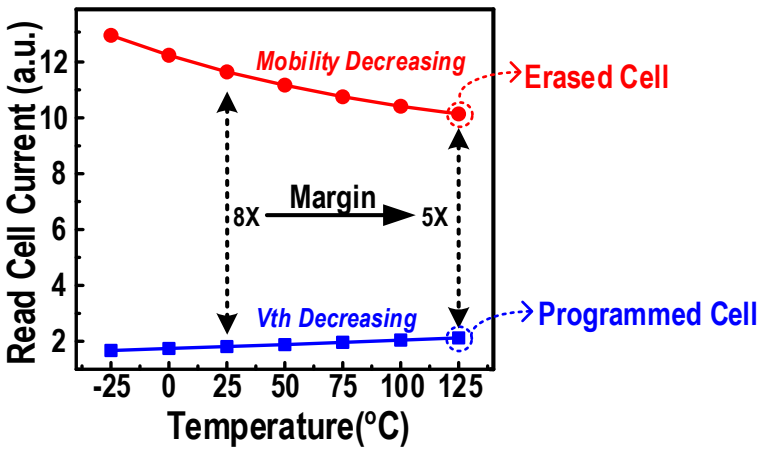


Figure 4.33 Simulated read cell current across temperature.

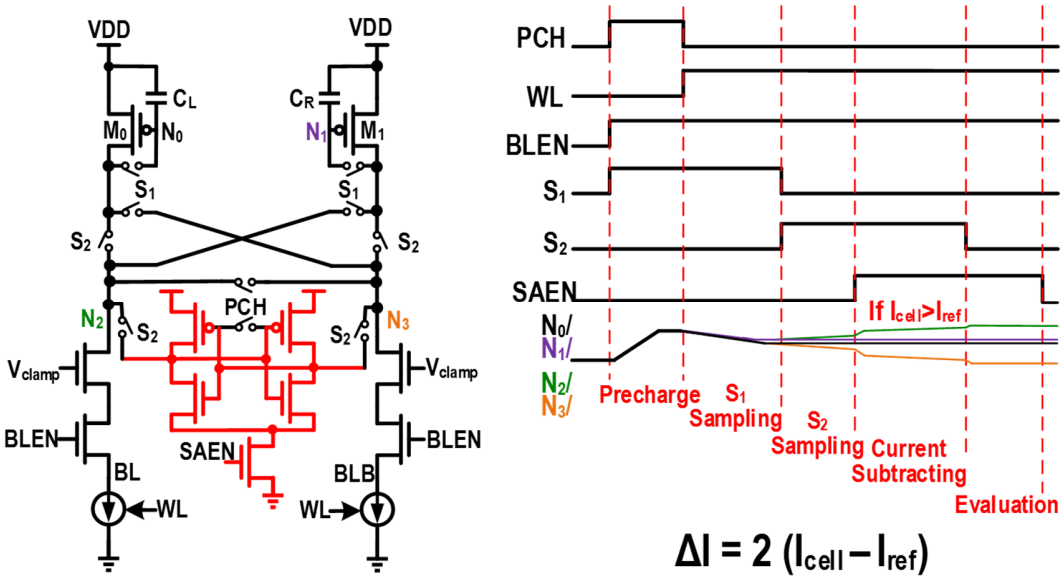


Figure 4.34 Circuit and timing diagram of margin-doubled current SA.

4.6.2 Margin-doubled Cross-Sampling Current Sense Amplifier Design

To address the degraded read margin, a margin-doubled cross-sampling sense amplifier is proposed, which requires 8 additional switches and 2 gate caps as shown in Figure 4.34. First, we turn-on S1 and BL enable to precharge both BL and reference BL. Then the WL is enabled. The cell current will flow in this path while the reference current will flow in the opposite. Currents are sampled by the gate caps, C_L and C_R. Now we turn-off S1 and turn-on S2. Therefore, reference current flows through the top left branch while cell current flows through the bottom left branch. Also the top right branch has the cell current but the bottom right branch has the reference current. As we use current sampling with the same transistors on top, mismatch of M0 and M1 can be alleviated. When we enable the latch-based second stage sense amplifier, the current is subtracted. The left node flows out Icell-Iref while the right node flows out Iref-Icell. Therefore, the current difference gets doubled, improving sense margin. Finally, we turn-off S2 and the second stage performs evaluation.

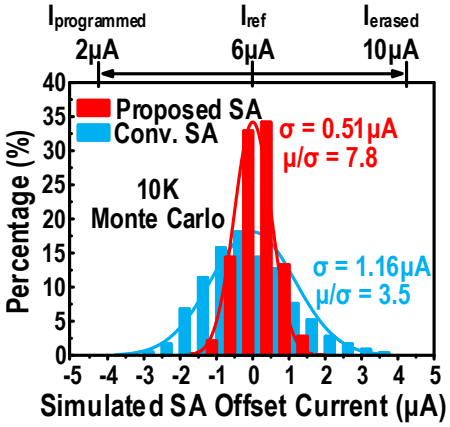


Figure 4.35 Simulated SA margin compared with conventional SA.

Because the doubled margin can be considered as doubled gain of the amplifier, the input offset is halved. According to simulation results in Figure 4.35, the offset of this sense amplifier

can be reduced by more than a factor of 2 due to doubled gain and alleviated mismatch of PMOS headers. Also, at high temperature, the simulated SA μ/σ is improved from 3 to 6.

4.7 Results

The flash macro is fabricated in 90nm embedded ESF3 NOR flash technology. A conventional compiled flash using the same bitcell is also fabricated for baseline comparison. Figure 4.36 shows the die photo of both.

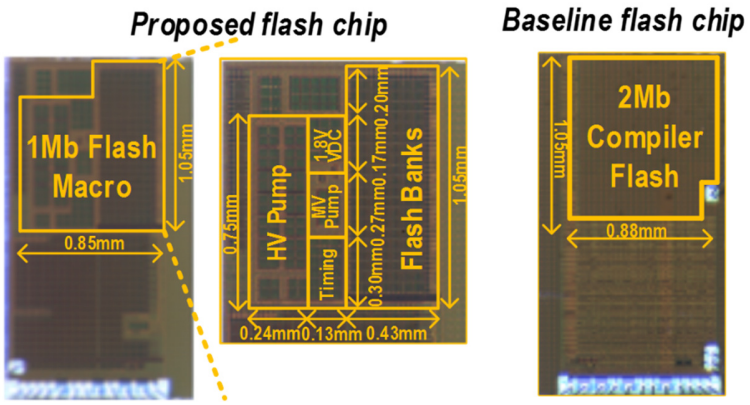


Figure 4.36 Die photos of the proposed flash chip and compiler baseline in the TSMC 90nm eFlash technology.

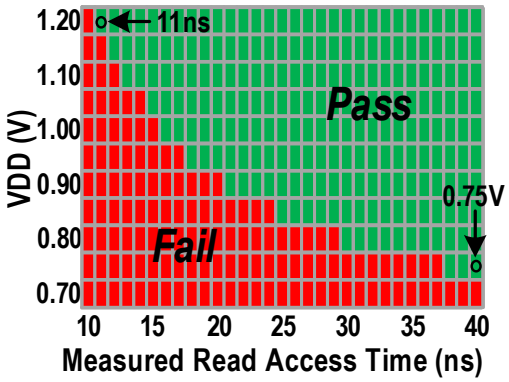


Figure 4.37 Measured Shmoo plot of flash read.

Figure 4.37 shows the measured Shmoo plot of the proposed flash macro achieving 11ns access time at 1.2V and 0.75V read VDDmin at room temperature. Measured average read VDDmin among 10 dies are 0.739V (Figure 4.38). Read VDDmin across −25°C to 125°C is shown in Figure 4.39, which also shows the program and erase power across temperature.

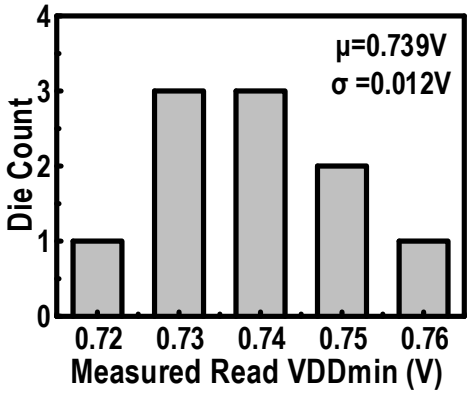


Figure 4.38 Measured read VDDmin distribution across 10 dies.

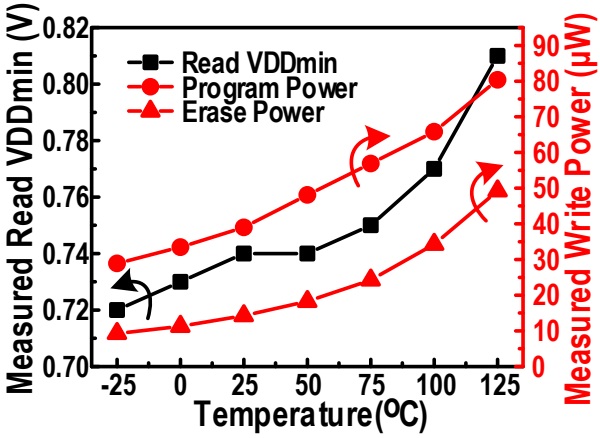


Figure 4.39 Measured read VDDmin, program power and erase power across temperature.

Measured erase and program powers at 25°C are 15μW and 39μW, respectively, which represent 242× and 87× reductions compared with the baseline (Figure 4.40). The erase and program energies are 9.4pJ/bit and 49pJ/bit, respectively, which are 30× and 22× lower than the

baseline. At 125°C, program (erase) power is 82µW (31µW), enabling reliable function of the battery-powered sensor system. At 11ns cycle time, read energy (power) is 2.2pJ/bit (1.618mW); at read cycle time suitable for sensor nodes (1µs), the design consumes 25µW and shows better power/frequency scaling than the baseline due to its lower leakage floor. Standby power is 5.4µW even with the charge pump regulation loops being active, which is a 4.5× reduction over the baseline. Table 4.3 compares the measurement results with baseline and other works.

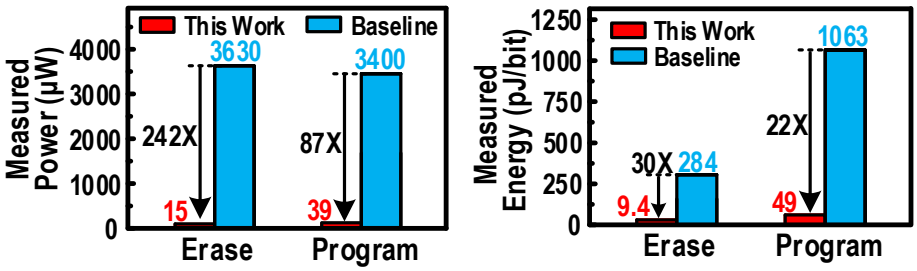


Figure 4.40 Measured power and energy comparison with baseline.

Table 4.3 Comparison table with baseline and other works.

Parameters		This Work		Baseline (Compiler Flash)		ISSCC16 [49]
Process (nm)		90		90		90
Cell Type		ESF3		ESF3		MONOS
Capacity		1Mb		2Mb		1Mb
Macro Area (mm²)		0.73		0.91		2.26
Power Supply Range (V)		2.2~2.8 / 0.75~1.4		1.6~3.6 / 0.95~1.4		3 ~ 3.6
Operation Unit		Erase:8kb Program:8b Read:8b		Erase:64kb Program:32b Read:32b		Erase:16kb Program:1kb Read:32b
Erase	Cycle (ms)	5		5		5
	Power (µW)	15(25°C)	49(125°C)	3630		NA
	Energy (pJ/bit)	9.4(25°C)	31(125°C)	284		NA
Program	Cycle (µs)	10		10		3000
	Power (µW)	39(25°C)	82(125°C)	3400		323(175°C)
	Energy (pJ/bit)	49(25°C)	92(125°C)	1063		946(175°C)
Read	Cycle (ns)	11(best)	1000	23(best)	1000	20
	Power (µW)	1618	25	5364	157	NA
	Energy (pJ/bit)	2.2	3.1	3.9	19.6	NA
	VCCmin(V)	0.75		0.95		NA
Standby Power (µW)		5.4		24.3		NA
Sleep Power (µW)		0.003		1.53		NA
Endurance		>10k		>10k		>100M
Retention		>10 years		>10 years		NA

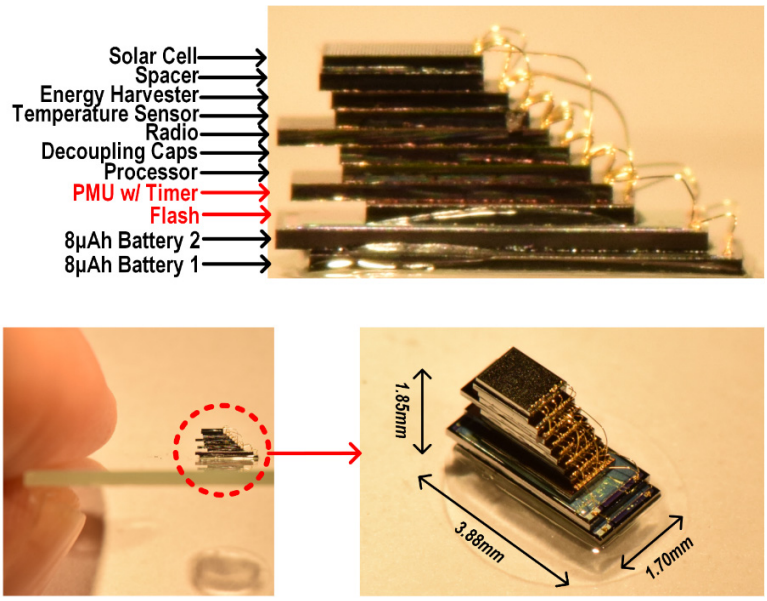


Figure 4.41 Photo of whole stacked system.

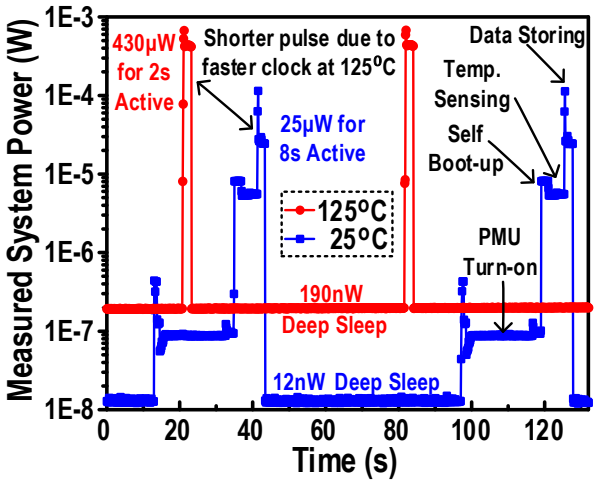


Figure 4.42 Measured results of system function power.

The flash macro is incorporated into a high temperature mm-scale sensor system that consists of multiple chip "layers", including two batteries, flash, PMU, processor, decoupling caps, radio, temperature sensor, energy harvester, and solar cell layer (Figure 4.41). Measured active system power drawn from the battery is 25µW and 430µW at 25°C and 125°C, respectively. Sleep

power is 12nW and 190nW at 25°C and 125°C, respectively, representing 3.3× and 63× reductions compared with a conventional system and greatly extending battery lifetime. The stacked system is fully functional at 125°C in stand-alone operation.

4.8 Conclusion

A battery-powered miniature sensor-node system with a 1Mb sub-100μW embedded NOR flash is presented for high-temperature sensing application. With NOR flash, the 3.88×1.70×1.85 mm3 stacked sensor-node system can be fully shut down with 190nW deep sleep power at 125^0C, achieving 63× improvement compared with conventional standby system power. The flash memory uses low-parasitic MIM caps in a combined Dickson and ladder pump topology to generate 13V with over 73% efficiency while maintaining reliability for the MIM caps. Ultra-low power voltage and current reference circuits are designed to further reduce power consumption. We introduce a sense amplifier that uses cross-sampling to double sensing margin, alleviating margin degradation of flash cell at high temperature and achieving 0.75V VDDmin. Measurements in a 90nm embedded flash technology show 30× and 22× lower program and erase energy, respectively compared with a standard flash macro. Flash program power is 39μW and 82μW at 25°C and 125°C, respectively, which enables battery-powered miniature sensor node systems to function reliably in various environments.

CHAPTER 5. STT-MRAM Design

5.1 STT-MRAM Concept

STT-MRAM is a promising candidate for next-generation non-volatile memory due to its excellent endurance, nano-second access time and great scalability [50]. Current with opposite direction can flip the spin polarity of free layer and therefore change the resistance of MTJ device. As MTJ device can be placed on top of CMOS transistors, no extra area is induced. Typically embedded MRAM uses 1T1R structure: one access transistor and one MTJ device.

Even though MRAM has demonstrated the potential as replacement for NOR flash, it still suffers from two issues. The first one is the limited read margin. As the TMR ratio of the STT-MRAM is only 200%, the resistance difference between the two states are very small, challenging sense amplifier design. Another concern is the write power. Hundreds of μA are required to flip the state of a single MRAM cell. Therefore, tremendous amount of power is consumed in write operation. To address these issues, both read and write assist methods are proposed.

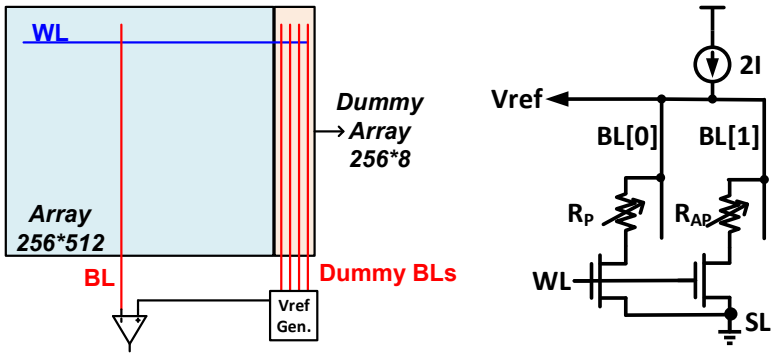


Figure 5.1 Read reference generation methods.

5.2 Proposed Read Assist

Since single-ended read has to be used due to 1T1R structure, read reference needs to be carefully designed to alleviate process variation. Figure 5.1 shows the proposed reference generation method. In each array, several dummy columns are added to generate the reference voltage which is shared by all sense amplifiers. When WL is enabled during read, two dummy cells in the dummy column will also be activated as shown in the right of Figure 5.1. One of the tow cells is R_P on the left and the other paralleled-connected cell is R_{AP} on the right. Once a reference current I_{ref} flows through this path, the output voltage will be $(R_P//R_{AP})*2I_{ref}$. PVT variation can be alleviated because the two reference cells are on the same row as selected cells tracking the PVT variation.

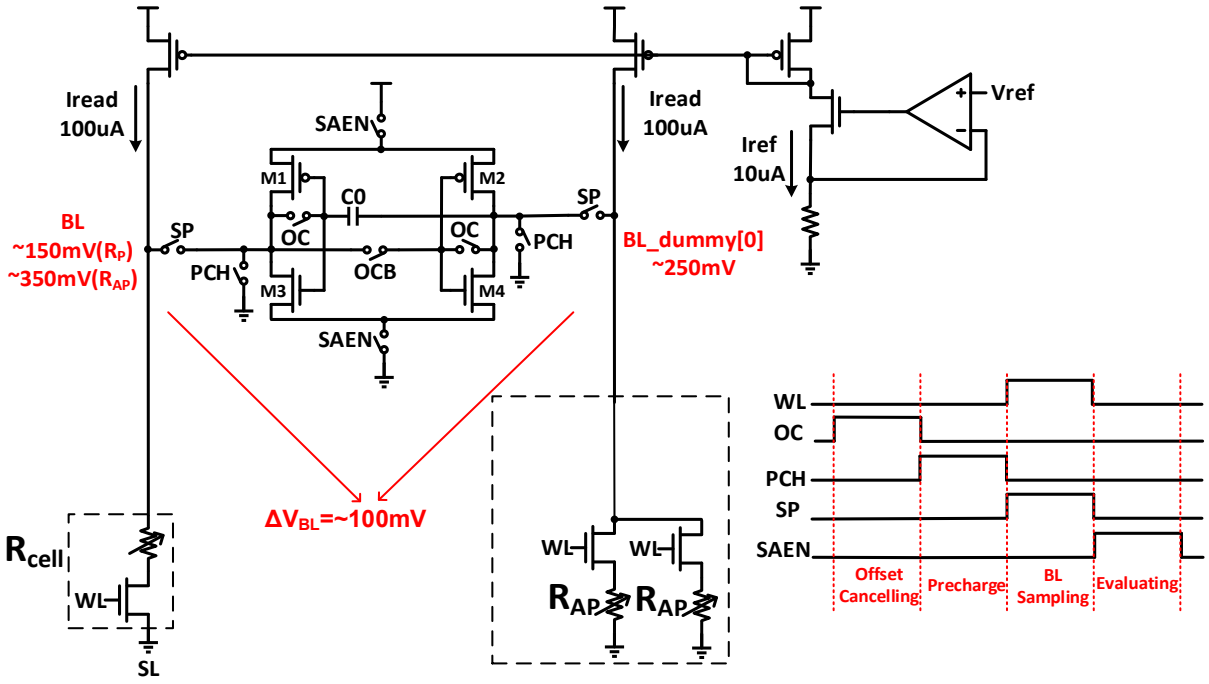


Figure 5.2 Read sense amplifier design.

We propose a constant current based sensing method using single-cap offset-cancelled sense amplifier as shown in Figure 5.2. Large current are applied on the BL to generate enough voltage

difference, in order to enlarge sensing margin and improve sensing speed. Constant current control is used to avoid read disturbance in this sensing mode. Then the large voltage difference can be quickly sensed with an offset-cancelled latch-based sense amplifier.

Several offset-cancellation methods [51-53] have been proposed to alleviate the mismatch of sense amplifier. However, they all use multiple caps with significant area overhead. A single-cap offset-cancellation method is proposed in this design as shown in Figure 5.3. Before precharge, an offset cancellation phase is inserted. In this phase, the inputs of the inverters will be connected to their outputs. The cap C0 will sample both trip voltages (V_L and V_R) of the inverters on the two sides. During precharge phase, the right side of C0 is discharged to 0. As a results, the left side stores V_L-V_R to mitigate the mismatch of the two inverters. BL sampling and Evaluation phases are same as conventional phases. As the V_L-V_R is always applied to the gate voltage of the left inverter in these phases, the offset is alleviated.

According to the simulation results (Figure 5.4), the input offset of the proposed sense amplifier is reduced by over 60% compared to the conventional sense amplifier with same transistor sizing. With approximately 1μm transistor width, the input offset of the proposed design is only 6mV.

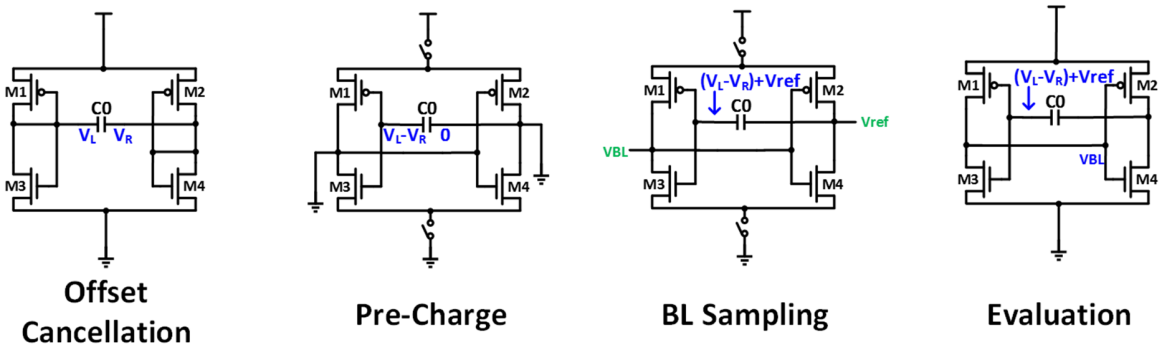


Figure 5.3 Offset-cancellation method.

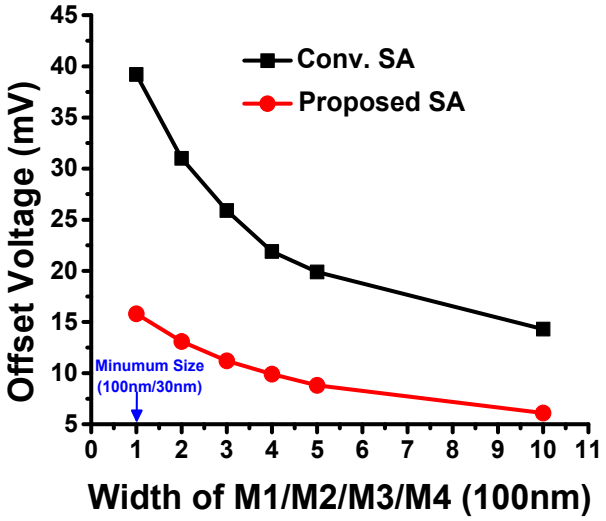


Figure 5.4 Simulation result of the input offset.

5.3 Proposed Write Assist

Figure 5.5 shows the proposed in-situ self-termination write method. During write, constant current (~300μA) is applied to the cell and the read sense amplifier will be reconfigured to detect write end. Once the write end gets sensed, the 'Stop' signal will be activated and disable the write driver on that BL. Each BL has its own self-termination control. BL voltage changes from high to low in both write 1 and write 0 case. Therefore, the detection method is same for both cases, simplifying the design of detection control.

This method can find redundant write and also auto terminate the write, saving power and improving reliability. The write detection circuit is reused from read sense amplifier with offset-cancellation, induce no area overhead. According to simulation results, the write power can be reduced by over 50%.

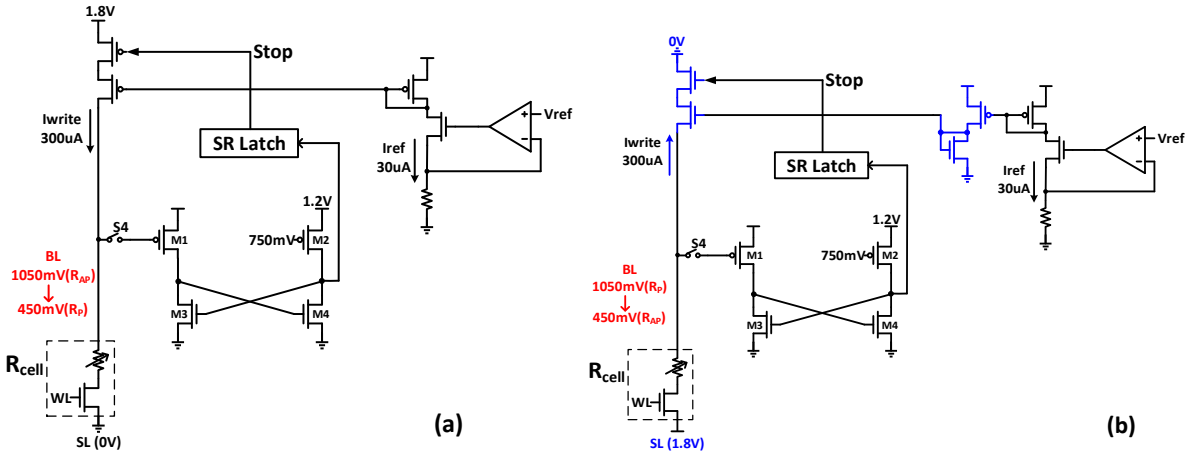


Figure 5.5 In-situ self-termination write 1 (a) and write 0 (b).

5.4 Results

The proposed 1Mb STT-MRAM was fabricated in TSMC 28nm technology. Figure 5.6 shows the die photo of the MRAM chip. The 1Mb MRAM macro occupies an area of 0.214mm^2.

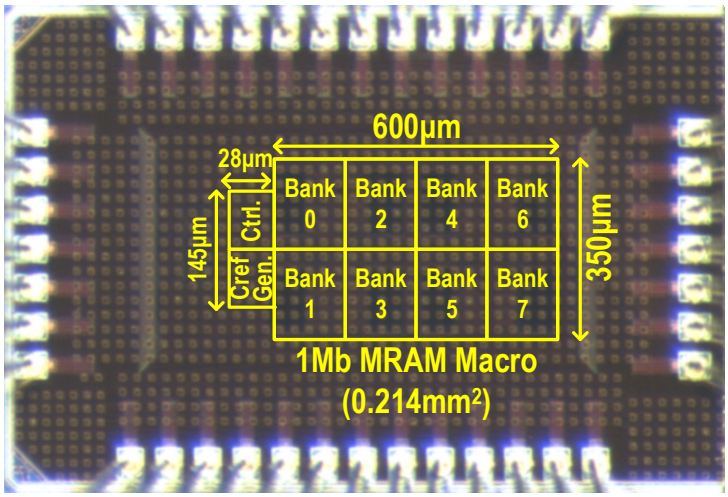


Figure 5.6 Die photo of the 1Mb MRAM Macro.

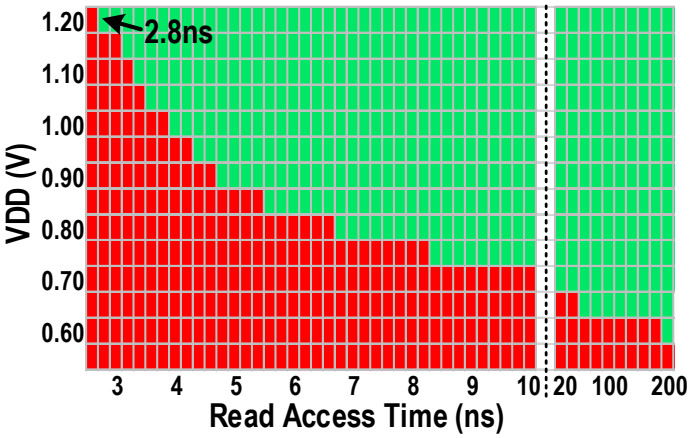


Figure 5.7 Measured shmoo plot of MRAM read operation.

Figure 5.7 shows the measured shmoo plot of the MRAM read operation. At 1.2V, the proposed design achieves 2.8ns read access time. Moreover, the read operation functions even at <0.6V power supply. Figure 5.8 shows the measured VDDmin across 10 dies for the proposed sense amplifier. Due to cancelled-offset, the sense amplifier works well at scaled power supply. The average VDDmin is 0.57V with a standard deviation of 0.019V.

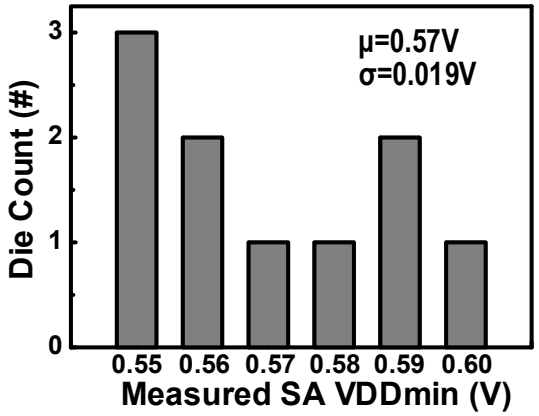


Figure 5.8 Measured VDDmin across 10 dies for the sense amplifiers.

As shown in Figure 5.9, the write fail rate decreases with longer write cycle time. To achieve 1E-5 error rate, the required access time has to be longer than 20ns. Since the write cycle time is very long, energy consumption is high as shown in Figure 5.10. With random data writing (50% data flip), the proposed self-write-termination can save energy consumption by 47% and 60% at 25°C and 120°C, respectively.

Table 5.1 compares this work with other MRAM works. This work achieves best read access time and power consumption with smallest macro area.

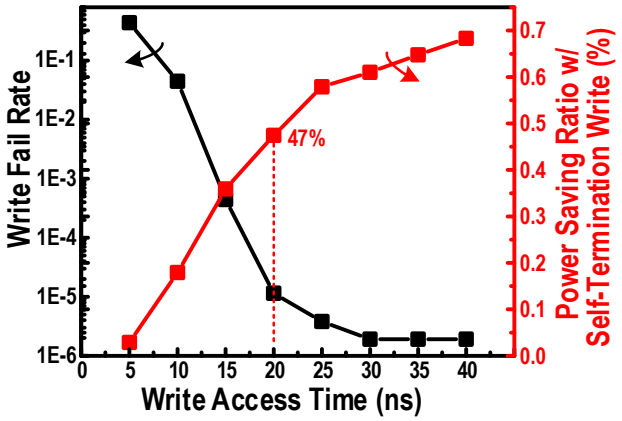


Figure 5.9 Measured write fail rate and power saving ratio across write access time.

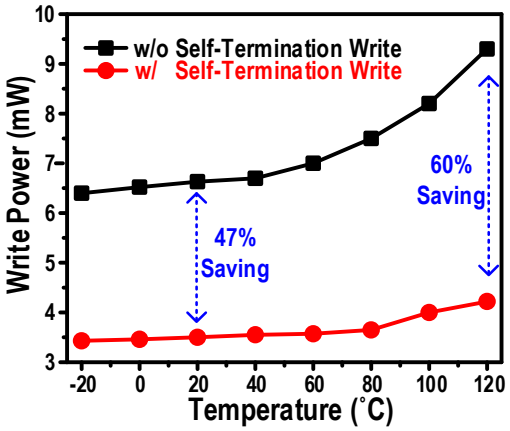


Figure 5.10 Measured power comparison between conventional write and self-write-termination write across temperature.

Table 5.1 Comparison with other MRAM works.

	This Work	ISSCC15 [54]	ISSCC13 [55]	VLSI12 [56]	ISSCC10 [57]	ISSCC09 [58]
Technology (nm)	**28**	65	40	90	65	90
Cell Type	**1T1MTJ**	2T2MTJ	1T1MTJ	4T2MTJ	1T1MTJ	2T1MTJ
Capacity	**1Mb**	1Mb	1Mb	1Mb	64Mb	32Mb
Macro Area (mm^2)	**0.214**	0.8196	0.57	3.54	47.124	91.02
Power Supply (V)	**1.2/1.8**	1.2/0.9/0.4	1.1/2.5	1.0	1.2	1.5
Word Length (bit)	**16**	256	32	32	16	32
Read Speed (ns)	**2.8 @25°C 3.6 @120°C**	3.3	10	8	30	12
Write Speed (ns)	**20**	3	NA	40	30	12
Read Power (mW)	**3.9**	21.6	NA	10.7	7.8	60
Write Power (mW)	**3.6**	55.4	NA	4.3	9.3	91

5.5 Conclusion

A low-power variation-tolerant 1Mb MRAM is proposed. For read assist, single-cap based offset-cancelled sense amplifier is proposed to improve sensing margin and self- reference generation method is used to track and mitigate the PVT variation during sensing. Moreover, in-situ self-write termination is applied to find redundant write and auto terminate the write in order to reduce write power by 47%. The chip is fabricated on 28nm embedded MRAM technology and achieves 2.8ns read access time and 0.57V VDDmin for sense amplifiers.

CHAPTER 6. Racetrack Converter for High-speed Imaging System

6.1 Introduction

Low-power and compact data converters are an essential part of sensor nodes as the link between the sensor and data processing. Also, in high-speed massive parallel sensors such as imagers, each photodiode includes a moderate-accuracy but compact analog-to-digital converter (ADC) for parallel data conversion [59]. CMOS implementations of such data converters face two challenges. The first is the difficulty of integrating ADCs with sensors in every pixel or channel due to the large area of analog circuits. This is exacerbated by poor scaling of analog circuits in CMOS due to process variation in advanced technologies [60]. The second is the high static power of analog data converters [61]. As a result, most high-speed image sensors only use column parallel ADCs in their sensor array to balance area/power and performance [62-63]. However image sensors with much higher frame rates are in demand for emerging imaging applications such as integral machine vision, time-of-flight (TOF) imaging, and three-dimensional high-definition television (3D-HDTV).

Recently, current-induced domain wall (DW) motion has driven the invention of spintronic devices that hold promise for non-volatility, high density, and low power. With perpendicular magnetic anisotropy (PMA) structures, hundreds of magnetic domains separated by DWs can be maintained in one nanowire for multi-bit non-volatile memory [16-18].

A novel spintronic-based data converter is presented that leverages the non-volatility, low power, and high density of spintronic devices. An n-bit racetrack ADC structure is

proposed using n magnetic nanowires with different configuration granularity for each bit [64]. The current-steering DW motion can convert n bits binary data or gray code in parallel. Exploiting the non-volatility of racetrack memory, the converted data is stored intrinsically, eliminating the need for additional memory cells and saving time spent writing data. Since most components are spintronic devices, the design achieves compact area, scalability, low static power, and no leakage. Compared to a conventional low power CMOS SAR ADC, the proposed racetrack ADC can achieve 1000× smaller area with comparable energy efficiency figure-of-merit (FOM). Also one potential application of the racetrack ADC is discussed: a high-speed imaging system using the 8b racetrack ADC as an in-pixel ADC. Results indicate that frame rate is increased by 50× compared to a CMOS digital pixel sensor (DPS) while retaining high fill factors as in analog pixel sensors (APS).

6.2 Racetrack Memory Device

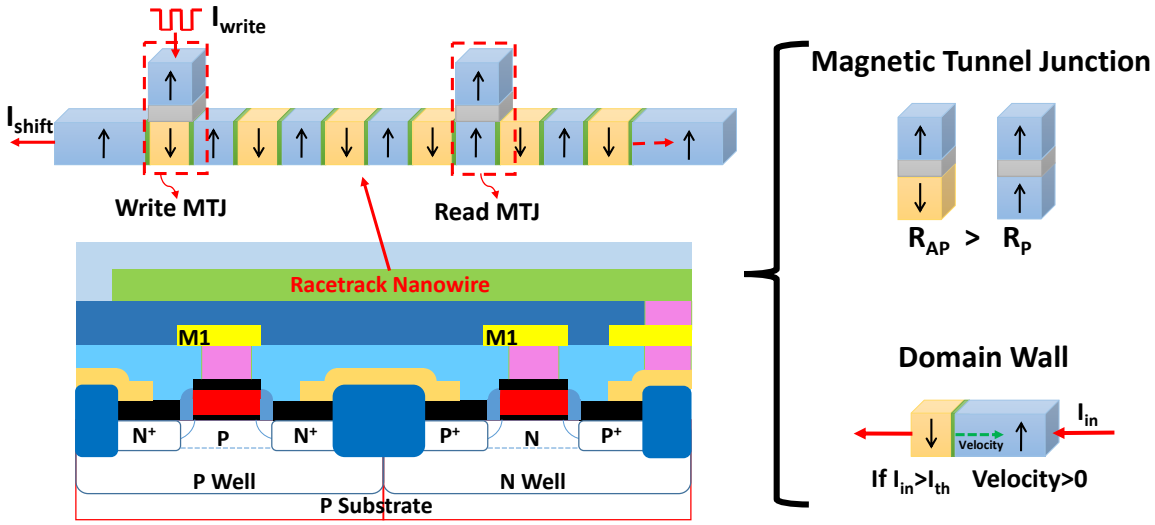
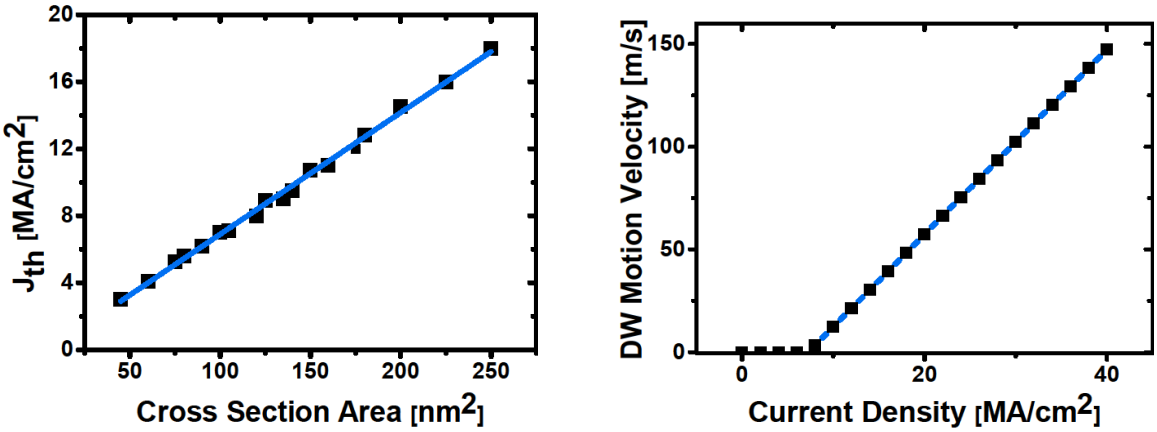


Figure 6.1 Structure of a racetrack memory device consisting of a magnetic nanowire and two MTJ heads as the read and write ports. The racetrack nanowire is manufactured on top of MOSFETs, avoiding planar area overhead.



(a) (b)

Figure 6.2 (a) Threshold current density decreases with reduced cross-sectional area; (b) Once current density exceeds the threshold, DW motion velocity linearly increases with higher current density.

The racetrack memory device is a magnetic nanowire comprising multiple magnet domains separated by DWs [18-21]. A single data bit is stored as the local spin polarity within the DW magnet strip at a given position. DWs can be shifted along the magnetic strip by induced horizontal charge current. Figure 6.1 shows the structure of a PMA racetrack memory consisting of one magnetic nanowire and two MTJ heads as the read and write ports. Given a current pulse I_{write} on the write MTJ, the magnetic domain beneath that MTJ in the magnetic nanowire will be nucleated through spin-transfer torque. At the same time, the horizontal shift current I_{shift} can move the data along the magnetic stripe. By alternatively asserting write current and shift current, the racetrack can store a sequence of DWs. Previous works have explored the potential of building hundreds of DWs in one magnetic nanowire [18, 21]. With such high density, the area efficiency can be as high as 1 F^2/bit [20], providing much higher density than other non-volatile memory technologies. The polarization of the magnetic domain beneath the read MTJ can be detected by sensing the resistance, which is affected through the tunnel magnetoresistance (TMR) effect. Reported MTJ reading access times for a megabyte-scale array are as fast as 4ns [24].

Moreover, this device can be implemented above CMOS transistors in the back-end process, reducing total area and interconnection delay.

Spin-dependent electron scattering can cause the charge current through the magnetic nanowire to be spin-polarized. When a spin-polarized electron crosses a DW, its spin-polarization will rotate 180 degrees from one magnetic domain to the other. To maintain total spin-angular momentum, the change of the current spin-polarization will be transferred to the local magnetization and create a spin-torque, causing the DW to move [18]. The DW moves along the flow of spin-polarized electrons, which is opposite to the direction of charge current. Both theoretical and experimental studies [61, 65-67] have shown that the threshold charge current density for DW motion in a PMA nanowire depends on the nanowire cross-sectional area. According to the adiabatic spin transfer torque model, threshold current density decreases with reduced width and thickness as shown in Figure 6.2(a) [61, 65-67]. When driving current exceeds the threshold current, the DW moves along the nanowire. Higher driving current can generate higher DW motion velocity. Using the compact model in [20], DW velocity can be described as

$$\boldsymbol{v = \frac{\beta \mu P}{\alpha e M_s}(J - J_{th})} \qquad \textbf{(6.1)}$$

where β is the non-adiabatic coefficient, μ is the Bohr magneton, P is spin polarization percentage of the tunnel current, α is the damping constant, e is the elementary charge, M_s is the demagnetization field, J is current density, and J_{th} is threshold current density. Velocity v can be increased with higher current density (Figure 6.2(b)). In a certain current range, the relationship between v and J would be quite linear [16, 17, 23]. Given this linear characteristic, current-induced DW motion is suitable for analog computation and data conversion in particular.

6.3 Propose Racetrack Converter

6.3.1 Overview of Racetrack Converter Operations

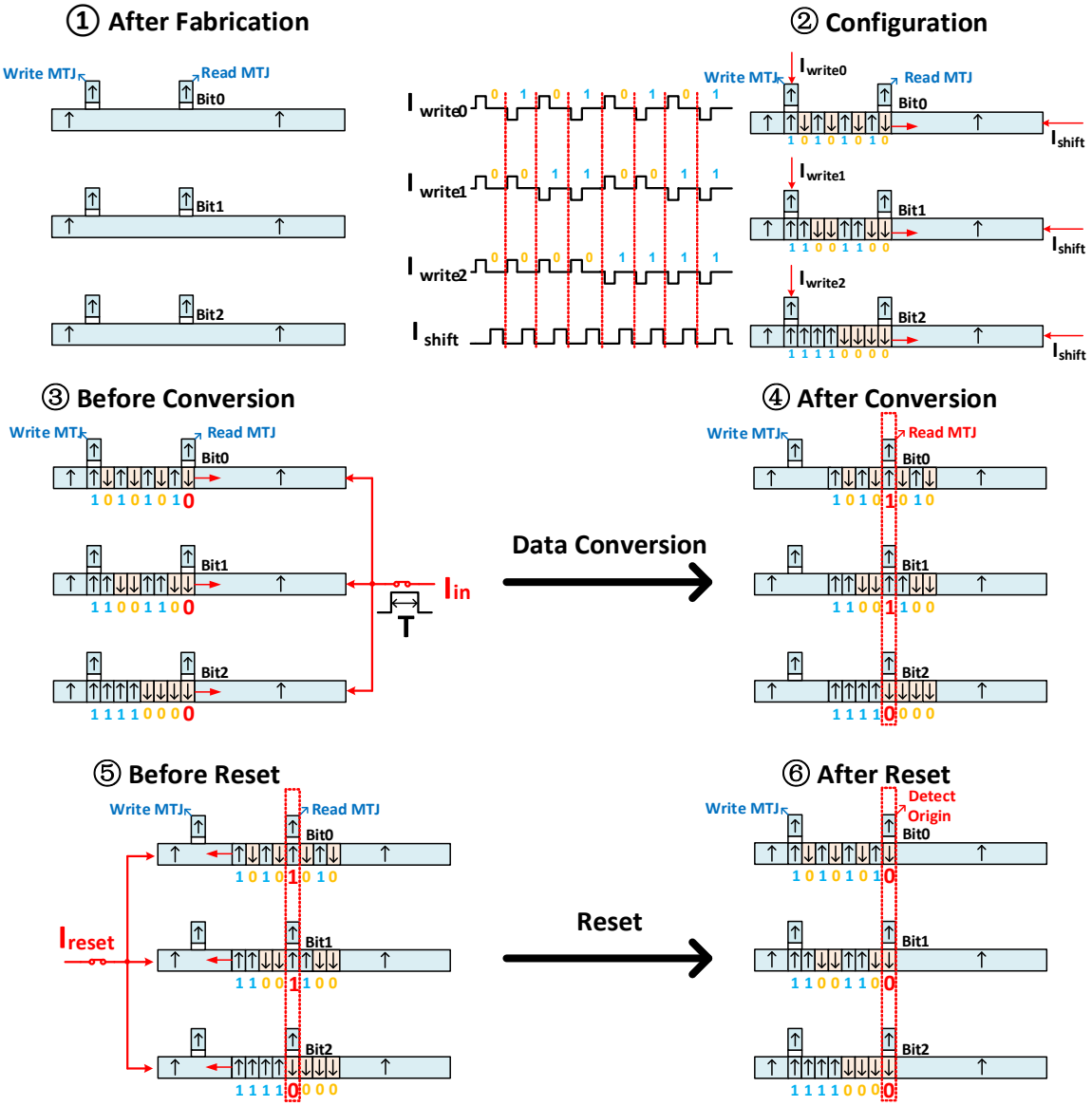


Figure 6.3 3-bit racetrack converter consists of 3 magnetic nanowires. After fabrication, each nanowire is configured with different DW granularity and represents an individual bit by current injection. During data conversion, current under test will flow through the nanowire, and all domain walls will move together. The moving distance is lineally proportionally to the current under test. After conversion, the converter need to be reset for next cycle.

1) Configuration: Figure 6.3 shows the structure of the proposed racetrack converter with 3 bits as an example. An n-bit converter requires n nanowires (Figure 6.4). Each nanowire will have 2^n DWs, one read MTJ and one write MTJ. Each nanowire will be configured differently such that each generates a single bit, from LSB to MSB. Then, the polarization of magnetic domain beneath n read MTJs represents the digitalized value from 0 to 2^n-1. This configuration is done only once post-fabrication, using the write MTJ port. When applying a positive current pulse on the write MTJ, the magnetic domain beneath that MTJ becomes spin-polarized with downward direction representing data 0; a negative current pulse generates upward spin-polarized magnetic domain (data 1). With a sequence of alternating write and shift current pulses, corresponding data can be stored on nanowires one by one. Altogether, such a design can store 256*8 bits on 8 magnetic nanowires to form an 8-bit racetrack data converter. According to [18, 68], domain wall write pulse is about 10ns using 1.2×108 A/cm^2 vertical current. Therefore, the write energy is less than 1pJ/bit [69]. As we only write once, the write power and latency is not important in this application.

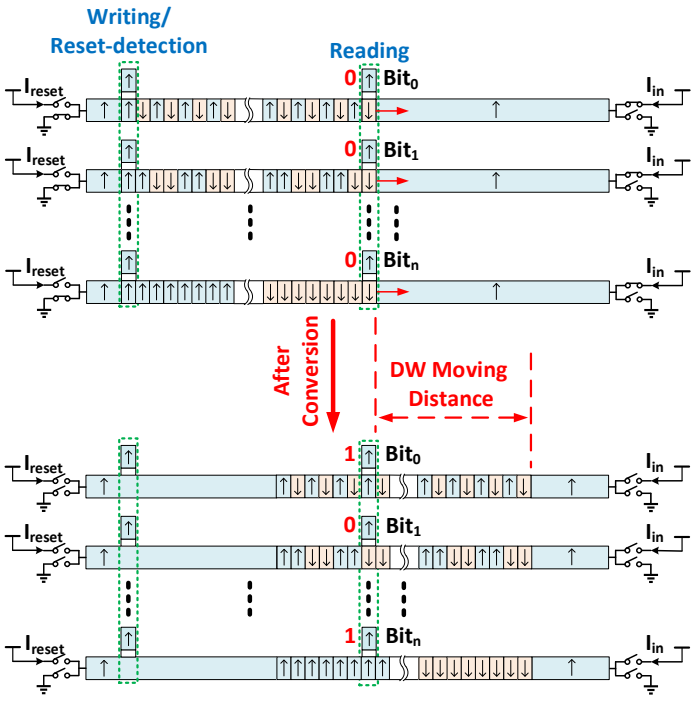


Figure 6.4 Data conversion scheme for an n-bit racetrack converter.

2) Data Conversion: As shown in Figure 6.3, the input current under measurement flows through the nanowire in the opposite direction of the reset current. In this case, all DWs move right simultaneously. As the current under measurement for each nanowire has the same value, the DWs in different nanowires move at the same velocity. After a fixed time *T*, the DWs will stop. The distance *X* that DWs move can be expressed as:

$$\boldsymbol{X = v * T = \frac{\beta \mu P}{\alpha e M_s}\left(\frac{I}{Area} - J_{th}\right) * T} \qquad \textbf{(6.2)}$$

The distance *X* is linearly proportional to the current *I* or the time *T*, which makes racetrack nanowires promising for both current-digital and time-digital converters.

3) Read: The polarization of the magnetic domain beneath the read MTJs stores the digitized value of the distance a DW has moved (ranging from 0 to 2^n-1). As the read MTJ head is upward spin-polarized, MTJ resistance with a downward spin-polarized nanowire domain could be 2-3× higher than that with upward polarized nanowire domain. Therefore, by sensing resistance of the read MTJ head above a given nanowire, a 0 or 1 state can be defined. Using sense amplifiers, the data can be read out as a digital value. In a more simplified design, write, reset and read can be performed with a single universal MTJ in the position of the read MTJ in Figure 6.3. The writing pulse in that case is applied to the single MTJ to configure the nanowires. With the input shift current flowing in the opposite direction, the data is left shifted one by one, rather than right. Reset detection is then performed by sensing all 0s as the beginning point using the universal MTJ.

4) Reset: After conversion, the write MTJ will subsequently perform reset point detection. Horizontal reset current flows through the nanowire to cause DW motion. When all DWs move back to their original position, write MTJ resistances undergo their resistance transitions, which is detected by sense amplifiers. Once the resistance change is sensed, the shift current is cut-off, signifying the completion of the reset phase. Sensing current is much smaller than writing

current and threshold current for DW motion, and therefore does not induce any change in the nanowire. If the latency of the sense amplifier is too long, over-reset or under-reset might happen. To address this problem, two-step reset should be used: 1) employing high current to quickly shift all DWs back with sense amplifiers coarsely detecting all 0s; 2) using small current to move each nanowire slowly with each sense amplifier finely verifying the reset point. After the first step, most of the nanowires will be reset to the original position, but some might have one-bit ahead or behind. As each nanowire has its own sense amplifier and fine reset control, the second step can carefully move each nanowire back to origin separately. The whole verification process can take less than 10ns which is only 20% of the whole cycle time (50ns). Both global coarse reset control and local fine reset control are simple logic gates (Figure 6.9).

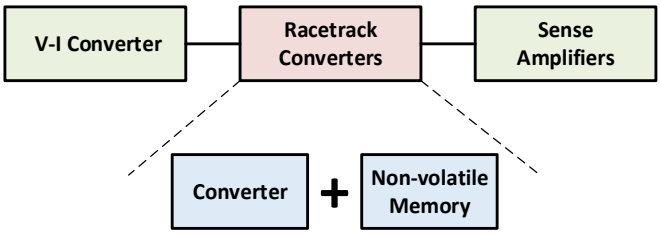
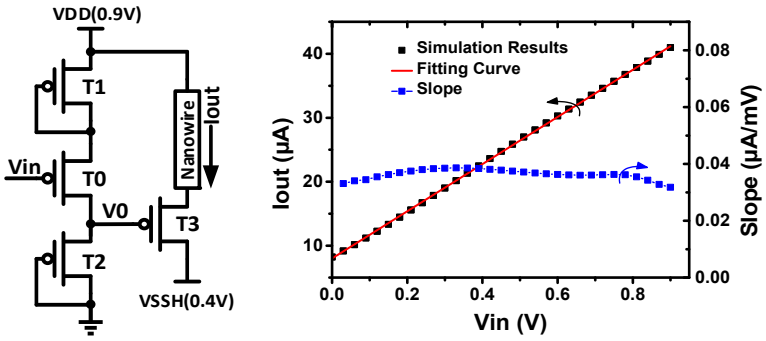


Figure 6.5 Racetrack converters function similarly to a combination of data converter and non-volatile memory.



(a) (b)

Figure 6.6 (a) Schematic of 4T all-PMOS V-I converter; (b) simulation results of its Iout-Vin characteristics.

6.3.2 Racetrack ADC

Most spintronic devices are suitable for current-mode computation because their characteristics have a direct mathematical relationship with current. The operation of mLogic [70], all-spin-logic [71], and domain wall neuron [61, 22, 24] are all based on current. The proposed racetrack converter is fully compatible with these current-mode spintronic logic devices. By combining with these other approaches, more complex current-steering mixed-signal systems can be implemented. However, most CMOS modules remain voltage-based. To realize integration, interfaces between CMOS and racetrack converter are needed. As shown in Figure 6.5, the racetrack converter includes the functionality of both a data converter and non-volatile memory. The interface circuits with CMOS need to provide current at the input of the converter and sense current at its output. Thus, the interface circuits mainly include sense amplifiers and a voltage-current (V-I) converter for the ADC.

For the ADC, the front-end interface should provide a current linearly dependent on input voltage. Figure 6.6(a) shows a 4T all-PMOS V-I converter with the racetrack nanowire as the load. T0, T1, and T2 in the first stage make up an attenuator (amplifier with gain less than 1). The output voltage of attenuator V0 will linearly follow the change of the input voltage Vin with opposite phase. The range of V0 is smaller than Vin, forcing T3 to operate in the velocity saturation region. As the electrical characteristics of a racetrack nanowire mimic a resistor, the current through the nanowire also changes linearly with input voltage. Moreover, the lower limit of current range is not 0 as T3 operates in the velocity saturation region across the full input voltage range. This lowest current can be designed to compensate the threshold current of DW motion. Figure 6.6(b) shows simulation results of the V-I converter's Iout-Vin characteristics. The transconductance of this 4T V-I converter is quite linear, with an adjusted R-Square value of 0.99996. Furthermore, this V-I converter is built using all PMOS, which increases its tolerance to

process corners. In addition, as a source follower the current is insensitive to the ground voltage of T3. Therefore, VSSH can be raised to achieve lower power.

The racetrack nanowire itself is a type of non-volatile memory. After conversion, data can be stored immediately in the non-volatile racetrack memory without area and timing overhead. With traditional CMOS current sense amplifiers (Figure 6.7(c)), stored data can be accessed. There is one situation to carefully consider. As shown in Figure 6.7(a), when a DW moves to the midpoint beneath the read MTJ, the resistance difference between the read MTJ and reference MTJ (with average resistance value) becomes very small. This may cause meta-stability in the sense amplifier and induce errors. Further, if most bits are approaching their flipping point, the error becomes significant. To avoid multi-bit flipping, gray code is used instead of binary code to ensure only one bit changes at a time (Figure 6.7(b)). Figure 6.7(d) shows the current offset distribution of the CMOS current sense amplifier with 10K Monte Carlo simulation. The standard deviation of the current sense amplifier offset is 0.7μA, much smaller than the sensing current range (from 20μA to 50μA).

Moreover, the self-reference sensing scheme is exploited to narrow the meta-stability region. As shown in Figure 6.8(a), an additional (reference) MTJ is placed next to the read MTJ. This additional MTJ serves as a reference with opposite phase to the read MTJ. Use the LSB gray code nanowire as an example. If the reference MTJ is placed 2 units distance away from the read MTJ on the top of the same nanowire, the DW polarity beneath the reference MTJ will always be complementary to that of the read MTJ. In the LSB gray code nanowire, bit 2 keeps commentary to bit 0(2-2) or bit 4(2+2). If the read MTJ resistance is high, that of the reference MTJ is low. When the resistance of the read MTJ approaches its middle value, the resistance of the reference MTJ similarly approaches its middle value from the opposite direction. As illustrated in Figure 6.8(b), this technique shortens the meta-stability window of the sense amplifier by 50%. For other bits in 3-bit gray code, 4 unit distance will keep results always complementary. Neither Gray coding nor self-sensing induce area or delay overheads.

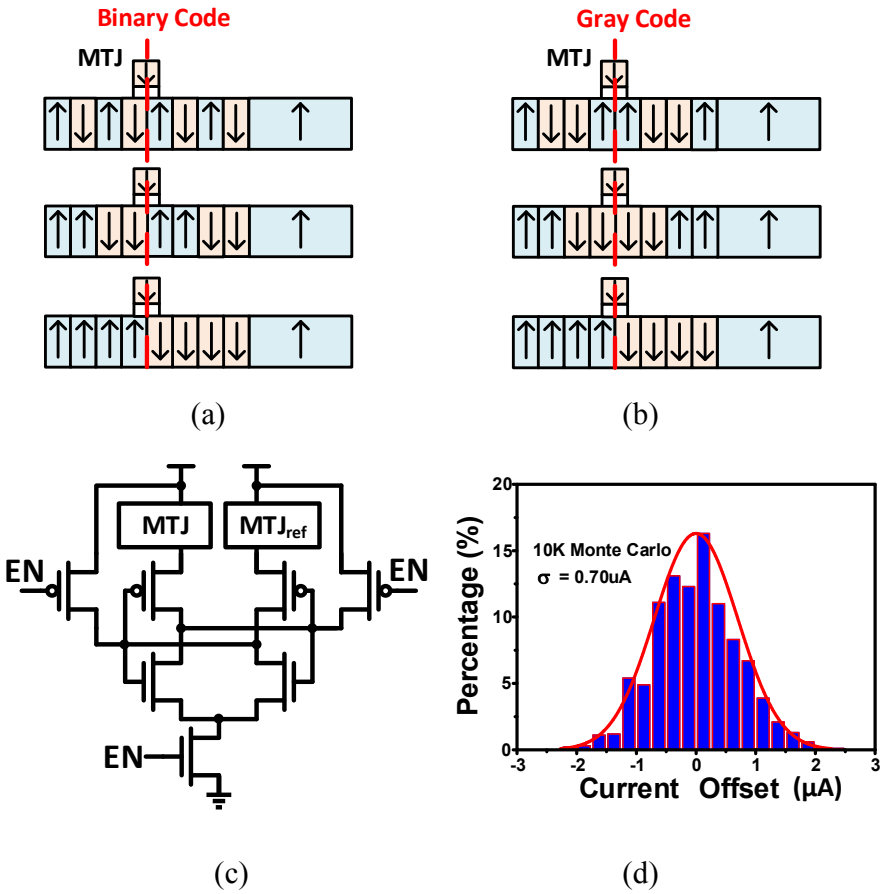
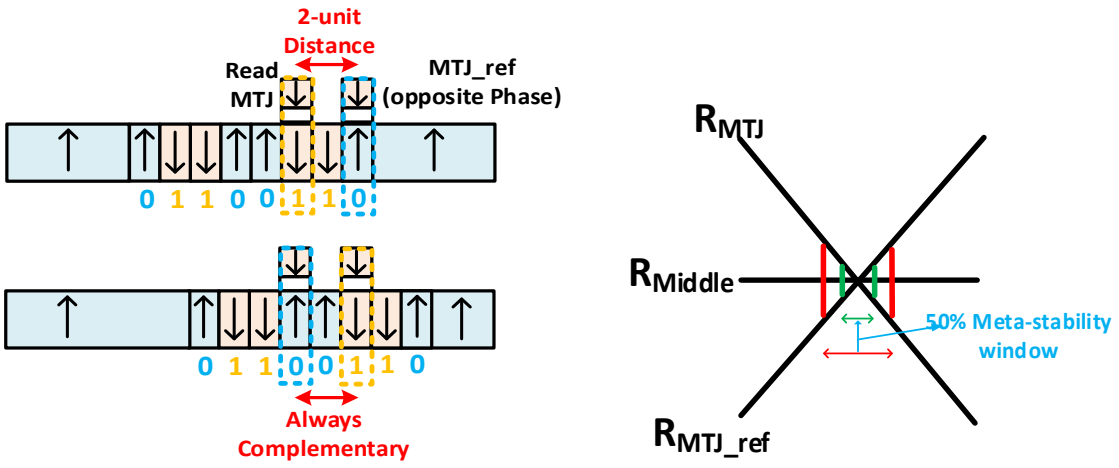


Fig. 6.7 (a) Midpoint meta-stability problem; (b) Solution with Gray Coding (c) Current sense amplifier schematic; (d) Sense amplifier current offset simulation results.

Figure 6.9 shows the block diagram of an 8-bit racetrack ADC including shared V-I converter, 8 racetrack nanowires and 8 sense amplifiers. Write, shift and reset switches are also shown in the figure. To minimize the mismatch among these nanowires, an offset compensation circuit is included in each nanowire. The main idea of the method is to add a tunable resistor connected in series with each nanowire. The simplest implementation of this tunable resistor is ratioed linear-region transistors connected in parallel. The number of connected on-state transistors can be tuned digitally according to the threshold current mismatch of the nanowire. Calibration is required to determine the value of tuning bits.



(a) (b)

Figure 6.8 (a) Self-reference Sensing: Using the LSB gray code nanowire as an example. If the reference MTJ is placed 2 units distance away from the read MTJ, the DW polarity beneath the reference MTJ will always be complementary to that of the read MTJ. (b)Sensing dead zone can be narrowed by 2× using self-reference sensing.

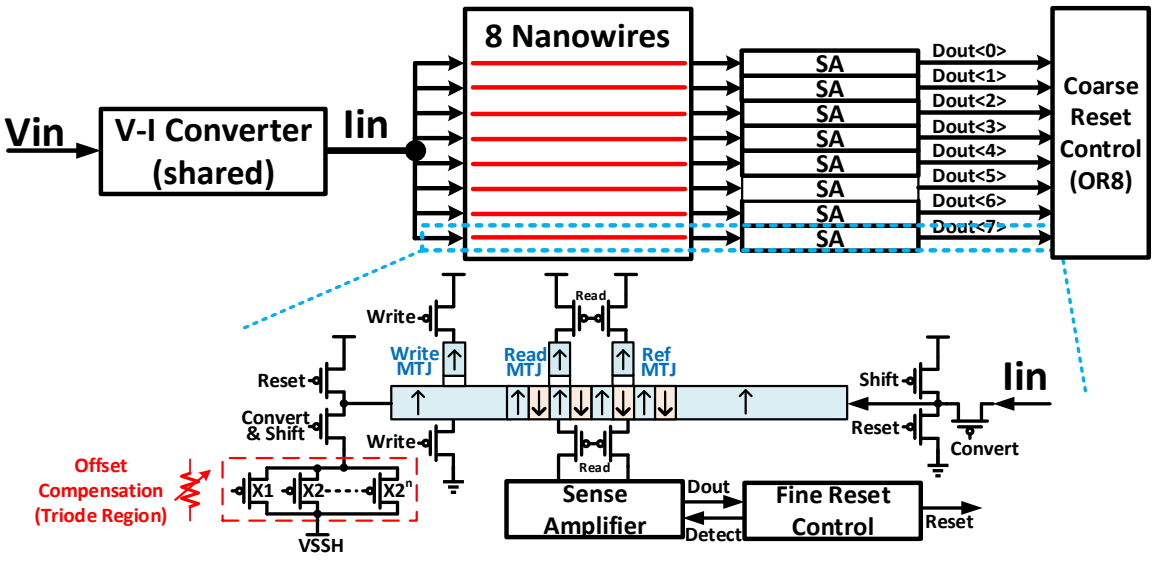


Figure 6.9 8-bit ADC block diagram and offset compensation method.

6.4 Uncertainty Analysis

6.4.1 Process Variation

Both CMOS process variation and racetrack nanowire variation will influence data conversion accuracy. For the CMOS part of the design, the 4T V-I converter is tolerant to systematic variation due to its all-PMOS implementation. In particular, systematic variation creates only a minor offset to the V-I conversion curve with little impact on slope and linearity. The offset can be cancelled using a simple N-well bias compensation method. However, the V-I converter remains sensitive to random variations induced by random dopant fluctuation (RDF) and line edge roughness (LER). The PMOS transistors are upsized to alleviate the influence of these random variations. In addition, random variation also leads to sense amplifier offset. Device sizing and/or auto-zero calibration techniques can be employed to enhance mismatch tolerance.

For the racetrack nanowire itself the major sources of variation include: 1) MTJ layer area; 2) tunneling oxide thickness; 3) cross-sectional area of the nanowire. Both 1) and 2) affect MTJ resistance [72] and may lead to read failure. The proposed self-reference sensing alleviate this influence. MTJ area can also impact the dynamic spin-polarization characteristic during configuration; this can be ameliorated by extending the write time and performing verification after configuration to ensure successful spin-polarization of each domain. Nanowire cross-sectional area can affect the threshold current density for DWs to move and shift the v-Jth curve of DWs. LER-induced cross-sectional area random variation is potentially the most severe variation for the proposed racetrack converter. Fortunately, similar to threshold voltage mismatch in CMOS devices, the threshold current mismatch arising from cross-sectional area random variation can be alleviated by variation-aware circuit design techniques (current offset compensation methods) or post-silicon calibration techniques. A mismatch compensation method is proposed as shown in Figure 6.9.

Advanced design techniques commonly used in CMOS converters, such as time-interleaving [73], can also be applied to improve its conversion accuracy or performance. Moreover, thanks to the extremely small area, an additional bit can be included to compensate for potential accuracy losses arising from variation with only 12.5% area and power overhead for 8-bit ADC, as an example. While adding one additional bit in CMOS ADC may even take about 100% area consumption because CMOS ADC area is in a quadratic relationship to the number of bits. But the racetrack ADC only requires one extra nanowire and one extra sense amplifier to add one bit. If the target is 8 bits ADC function, 9-bit racetrack ADC can be used to realize the 8-bit function with only 13% extra area overhead and the accuracy can be improved compared to using only 8-bit ADC.

6.4.2 Noise

The proposed racetrack converter works in current mode during data conversion. Compared with a conventional voltage mode CMOS converter, current mode computation suffers less from noise [74]. Moreover, during conversion a constant current flows through the nanowires without any switching. Therefore, the proposed racetrack converter is immune to switching related noise, making thermal noise the dominant noise source during data conversion. Both the V-I converter and racetrack nanowires will contribute thermal noise. Spintronic devices generate much less thermal noise than MOS transistors because of their smaller resistance [75]. Based on previous analysis, the total integrated thermal noise current of the converter could potentially be 3 orders of magnitude smaller than the input current. The simulated noise standard deviation (both thermal and flicker) of both V-I converter and nanowire is ~0.42mV.

6.4.3 Stability and Reliability

Using PMA magnetic material, the current-driven motion in a domain wall is not sensitive to pinning and local magnetic fields or temperature [16, 18]. 10-year retention at 150°C can be

achieved and endurances above 1010 cycles have been reported with 109A/cm^2 write current density in 90nm technology, which demonstrates the great reliability [17, 67, 76, 77]. Thanks to the low resistance of the nanowire (~10kΩ for 3.5nm×30nm×16μm [22]), the joule heating power is less than 10μW, comparable to CMOS.

This combination of high endurance and excellent retention indicates the technology is a good match for high sampling rate non-volatile data conversion. Moreover, as converted data will be sent to a processing module immediately after conversion, there are no concerns with thermal stability related retention.

6.5 Simulation Results and Analysis

Compact Verilog-A models of MTJ and racetrack nanowire are built based on published experimental data [16-17, 19-20]. The dimensions of each nanowire are designed to be 3.5nm×30nm×16μm, compatible with a 32nm technology. Co-simulation with CMOS circuits (a commercial 32nm SOI technology) is performed by SPICE simulator. The CMOS circuits and nanowires are stacked and their area are matched.

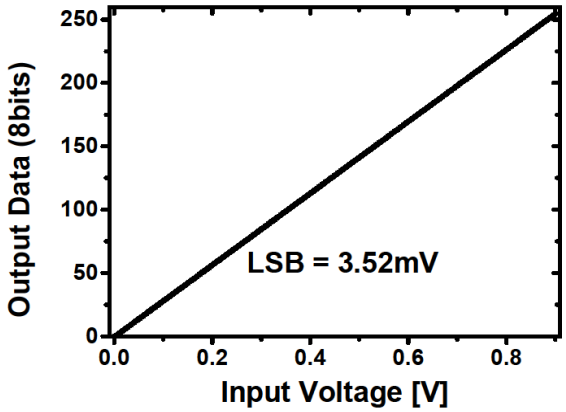


Figure 6.10 Simulated data conversion of the ADC.

Figure 6.10 shows data conversion simulation results of an 8b racetrack ADC. The ADC input voltage range is [0, 0.9V] with 3.52mV LSB. The ADC has a variable input voltage with fixed sampling time for DWs to move.

The V-I converter changes the ADC input voltage to a current for data conversion. The current range is 7-40μA (6.7-38 MA/cm^2) for this 32nm technology. The V-I converter can tune the output current range to guarantee the DW velocity increase linearly with higher current density. In state-of-the-art experimental results, the needed current density to move the DW by 250nm within 2ns is 18MA/cm^2 at 90nm technology [17, 76]. Moreover, [17] experimentally shows a quite linear relationship of DW velocity upon current density range from 18 MA/cm^2 to 50MA/cm^2.

As the threshold current density will decrease with reduced cross-sectional area [75-77], less than 6.7MA/cm^2 threshold current density for 32nm technology is achievable as mentioned in [61]. Also, good linearity of current density range from 6.7 MA/cm^2 to 38MA/cm^2 in 32nm technology should be achievable. The minimum DW velocity at 6.7MA/cm^2 will be 1m/s, with which one bit code (30nm) can be shifted within 20MHz cycle time.

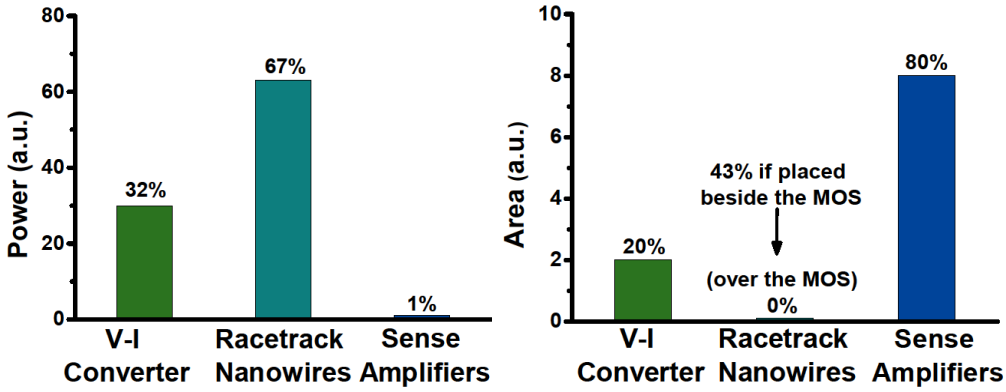


Figure 6.11 Power (a) and area (b) breakdown of each ADC component. Racetrack nanowires consume the most power though area overhead can be ameliorated by placement above MOSFETs.

Figure 6.11 shows the power and area breakdown of each component. Among the three parts, the racetrack nanowires dominate power consumption, taking more than half of the total power. The nanowires operate continuously during conversion and possess resistor-alike electrical characteristics.

Sense amplifiers only operate for a short time (less than 1ns) after conversion and remain in a low-power standby mode during conversion. Compared with sense amplifiers, the V-I converter consumes more power yet less area. By raising VSSH, V-I converter power can be lowered. As racetrack nanowires can be placed on the top of the MOS, they do not induce extra area overhead if carefully designed.

Table 6.1 shows the characteristics of a racetrack ADC in 32nm technology. At 20MHz, the total power is 96μW. Furthermore, the area is only 10μm^2, which is 3 orders of magnitude smaller than state-of-the-art CMOS ultra-low power SAR ADCs with comparable FOMs (~20 fJ/conversion-step). Racetrack ADC power consumption is input-dependent. As shown in Figure 6.12(a), higher input voltages generate higher currents through the nanowires and V-I converters, consuming higher power. The average power also increases with higher sample rate of the ADC (Figure 6.12(b)). However, because of the uncertain DW velocity linearity at ultra-high currents, a modest operating frequency of 20MHz is used to ensure reliability. With a wider linear region of DW velocity, higher sampling rates can be achieved.

Another advantage of the proposed racetrack ADC is that total area and power increase linearly with resolution rather than exponentially (Figure 6.12(c)). Adding one bit requires only one additional magnetic nanowire and one sense amplifier. Racetrack ADCs also benefit from the significant scalability of spintronic devices. With technology scaling [78, 79], the total nanowire length can be shortened, which will lower the required velocity to achieve the same sample rate. The cross-sectional area will be minimized, lowering the threshold current density of DW. As DW motion is based on current density, smaller cross-sectional area also translates to

smaller current needed to achieve the same velocity. Therefore, to realize a constant sampling rate, average power reduces cubically with scaling (Figure 6.12(d)).

There are 4 major factors affecting ADC non-linearity: 1) Non-linearity of the DW velocity upon current density; 2) Mismatch of the nanowire; 3) Sense amplifier offset; 4) Non-linearity of the V-I converter.

For 1), the non-linearity of the domain wall velocity upon current density can deteriorate the DNL. However, it's hard to estimate the accurate non-linearity with insufficient and sparse published experimental data. With compact model, it's ideally linear. For 2), nanowire mismatch will also shift the DNL. According to [80, 81], the standard deviation of domain wall mismatch can be approximately 5%. Fortunately, with the proposed mismatch-compensation method (Figure 6.9), the nanowire mismatch can be minimized to less than 2%, affecting DNL with 2% variance. For 3), the sense amplifier offset will shift each code randomly. Using Monte Carlo simulation, the standard deviation of the current offset is 0.7μA. With 0.1V across the read MTJ and the reference MTJ, the current under sensing ranges from 20μA to 50μA. Thanks to our self-reference technique, the effective sigma offset is 0.35μA which can shift DNL by 1.2% LSB. For 4), according to the simulated results, the non-linearity of V-I converter will contribute 5nA shift in average, which is about 4% LSB. In total, the standard deviation of DNL shift is 7.2% LSB.

Noise simulation is performed for both V-I converter and nanowires (as resistors). The standard deviation of the noise (both thermal and flicker) is 0.42mV which is 12% LSB. By using the method in [82], the SNDR is 46.21 with 12% LSB noise and 7.2% DNL shift. The ENOB is 7.38.

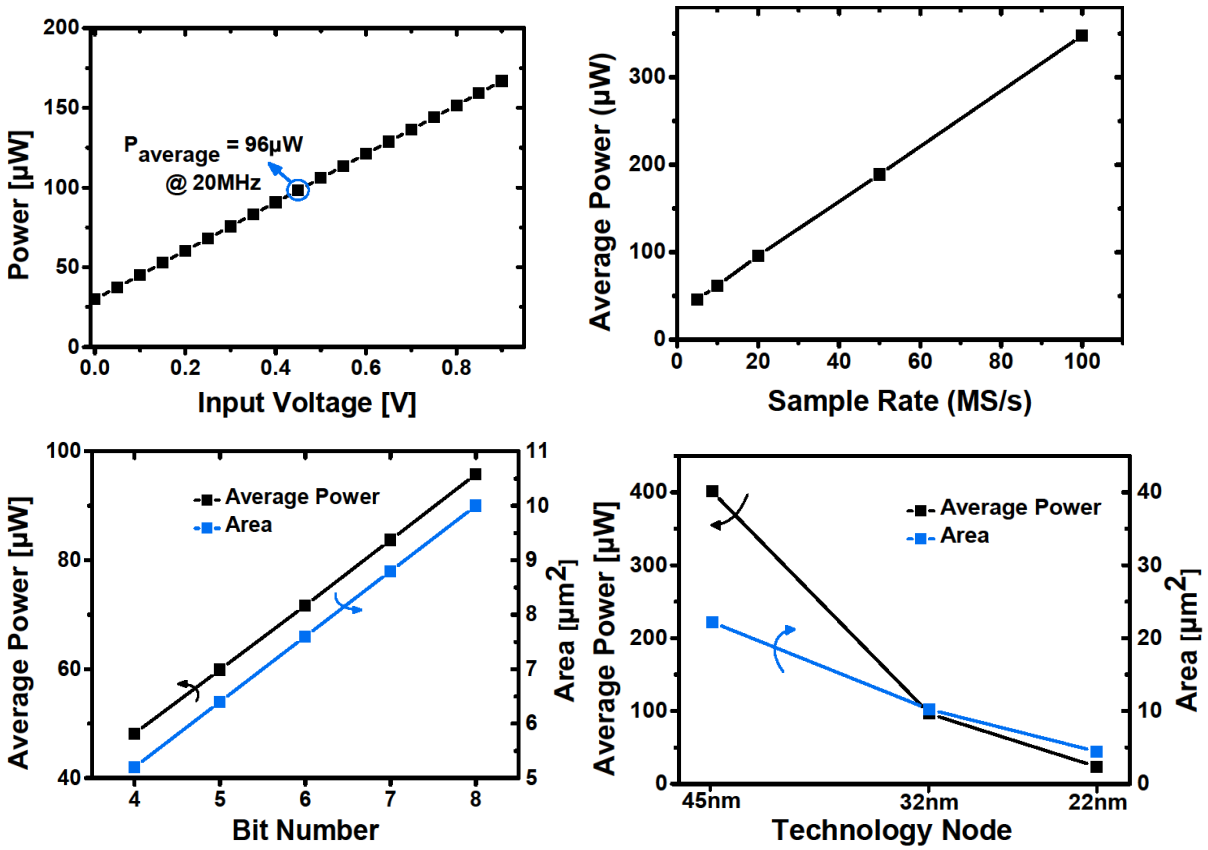


Figure 6.12 (a) Relationship between power and input voltage; (b) Average power increases with higher sample rate; (c) Total area and power increase linearly with more bits; (d) Average power reduces cubically with technology scaling.

Table 6.1 Comparison to recent 8b low power CMOS ADCs with comparable sampling rates

	CMOS ADC [83]	**CMOS ADC [84]**	**Racetrack ADC**
Technology (nm)	90	40	32
Sample Rate (MHz)	10	20	20
Resolution (b)	8	8	8
Power (µW)	26.3	84.9	96.5
FOM (fJ/conv.)	12	19.2	21
SNDR (dB)	48.50	44.90	46.21
ENOB	7.70	7.17	7.38
Area (mm^2)	**0.021**	**0.0153**	**0.00001**

6.6 High-Speed Image Sensor with Racetrack ADCs

High speed imaging systems employing in-pixel ADC, also known as digital pixel sensor (DPS) have several advantages over widely-used conventional analog pixel sensor (APS) architecture with column-wise ADC, including much higher speed, better scalability, and less noise (read-related column fixed-pattern noise and column readout noise). In particular, frame rate can be improved by more than 10× over APS with column-based ADC. With in-pixel ADC, only digital data is read out, which is faster and consumes lower power than reading analog values followed by conversion. However, the major bottleneck limiting the application of CMOS DPS is the large pixel size and low fill factor due to the area overhead of ADC and memory. Reference [59] reported a high-speed DPS with per-pixel single-slope moderate-accuracy ADC and 8b 3T DRAM. The dynamic range and frame rate are greatly enhanced with DPS architecture yet area and fill factor are unreasonably high [62-63].

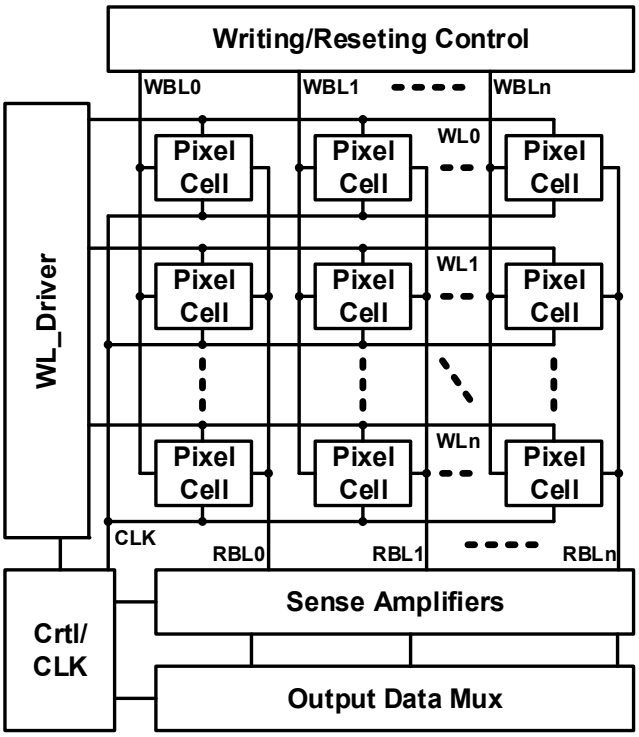


Figure 6.13 DPS block diagram implemented with racetrack ADC.

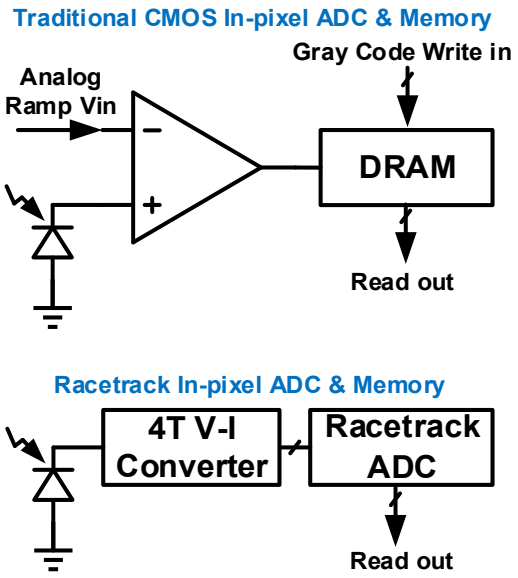


Figure 6.14 Digital pixel cell comparison. Unlike CMOS single-slope ADCs, the racetrack ADC does not require analog ramp voltage and write-in data. Readout can be done with shared sense amplifiers like memory readout.

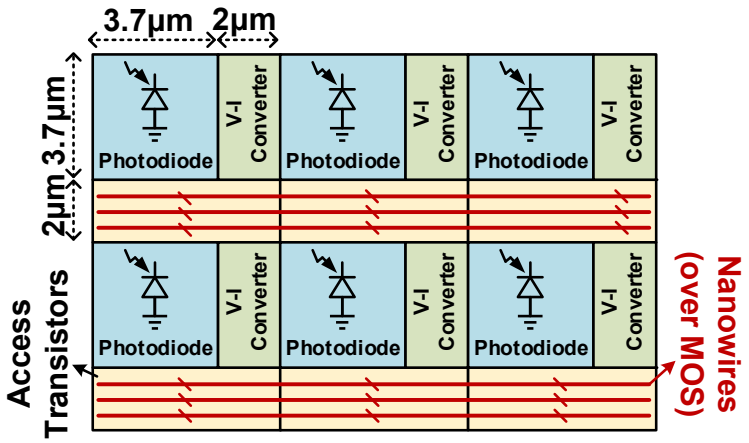


Figure 6.15 Layout implementation of 2×3 pixels. Racetrack nanowires can be placed on top of the access transistors.

The proposed racetrack ADC is a combination of ADC and non-volatile memory with extremely compact area, making it well suited to the DPS image architecture that commonly

relies on a CMOS moderate-accuracy ADC and separate memory. Figure 6.13 shows the DPS block diagram with the proposed racetrack ADC in each pixel. Unlike a CMOS single-slope ADC that requires analog ramp voltage from peripheral DAC and write-in data for memory, a racetrack ADC only requires CLK from a peripheral block during conversion, simplifying the required peripheral circuits and alleviating noise and I-R drop (Figure 6.14). The system structure is similar to a typical non-volatile memory bank. Sense amplifiers are placed at the bottom of the array (Figure 6.15) and shared by the array. Therefore, their area are not included in the pixel cell. Both read and configuration operations are performed row by row like a non-volatile memory. During read, when one row is accessed, the read MTJs will be connected to the shared sense amplifiers to read out the data. After one row read-out, the address will change and activate the other row alternately. After read of the whole array, all racetrack nanowires need to be reset for next conversion. Configuration is also performed row by row. When one row is activated, BLs will be connected to each write MTJs header through PMOS switches. Write current flow through BLs to the write MTJ and flip the polarity beneath it, and then shift current flow through the WL with PMOS switches. Conversion will be done inside each pixel all at the same time.

Both the V-I converter and access transistors can be implemented with only PMOS devices, further minimizing the required area as large N-P well spacing is not needed. Moreover, the racetrack nanowires can be placed on top of PMOS transistors. In this way, the nanowires will not induce extra area overhead. Figure 6.15 illustrates the layout implementation of 2×3 pixels. The nanowire is somewhat long but very narrow, hence 1×3 pixels can be arranged together to match nanowire length with 3×8 nanowires placed above the access PMOS. In this arrangement, area is still dictated by transistors rather than the nanowires. Thus, the fill factor is significantly improved over CMOS alone.

Table 6.2 compares CMOS APS and DPS image sensors with a Racetrack DPS image sensor. Sensor fill factor matches CMOS APS and is 3× better than CMOS DPS. Frame rate is improved

by 50× with lower power consumption. The proposed racetrack ADC is very promising for this high-speed imaging application.

Table 6.2 Comparison between CMOS APS, CMOS DPS and Racetrack DPS

	CMOS APS [62]	**CMOS DPS [59]**	**Racetrack DPS**
Sensor Fill Factor	40%	15%	42%
Frame Rate	3 500	10 000	>500 000
ADC Resolution (b)	12	8	8
ADC Conv. Time (ns)	500	25	50
Energy Per Frame (μJ)	280	5	0.04

6.7 Conclusion

The potential of current-induced domain wall motion has been explored for data converter application and an ADC design scheme is proposed based on racetrack magnetic nanowires. The 8-bit ADC can achieve 21 fJ/conversion-step, while the area is less than $10\mu m^2$, which is 1000× smaller than state-of-the-art CMOS ADC with similar energy efficiency. The results indicate that racetrack converters hold promise for future low-power small-area applications requiring multiple ADCs. A high-speed racetrack ADC-based DPS image sensor system is also proposed to show one potential application of this design.

CHAPTER 7. Neural Network with Spintronic Devices

7.1 Introduction

CMOS implementations of neural network suffer from large area for weight storage and high standby power. Spin devices such as magnetic tunnel junction (MTJ), domain wall (DW) and racetrack nanowires have demonstrated great potential for memory [17, 18], logic [70], and analog computation [23] due to their non-volatility, zero standby power and compact area, making them promising candidates as neural network components. There have been several recent works on spintronic neural networks. LSV neuron [85] only uses binary weights and suffers from the short diffusion distance associated with spin current. The DW binary-threshold neural network [22], combining with RRAM or PcRAM, requires more masks and complicates fabrication. DW synapse [86] has analog programmability but its spin current limits the number of synapses connected with neurons. None of them realized analog programmability with all charge current. Also, they are all feed-forward binary-threshold neural networks (BTNN), which are limited in their effectiveness in that more neurons and hidden layers are required to perform the same function compared with a rectified-linear neural network (RLNN). Time-related neural networks such as recurrent neural network (RNN) have not been explored yet.

A novel spin synapse device is proposed with analog programmability using all charge current. In this device, the resistance can be varied in an analog fashion according to the DW position, which can be moved by charge current injection instead of spin current diffusion. The proposed synapse devices are placed in a cross-bar array configuration to form a dense neural network. Using current summation on the bit-line, DOT product

function can be realized. Using ultra-compact racetrack converter [23], a more efficient RLNN can be built. A novel majority voting circuit is proposed in the final-decision layer for recognition tasks. The spin RLNN reduces area by 67% and energy by 69% compared to spin BTNN with equal error rates. Furthermore, integrating with time, current-induced DW motion can be used for time-related analog storage to form recurrent neuron.

7.2 Components of Spin Neural Network

7.2.1 Spin Synapse

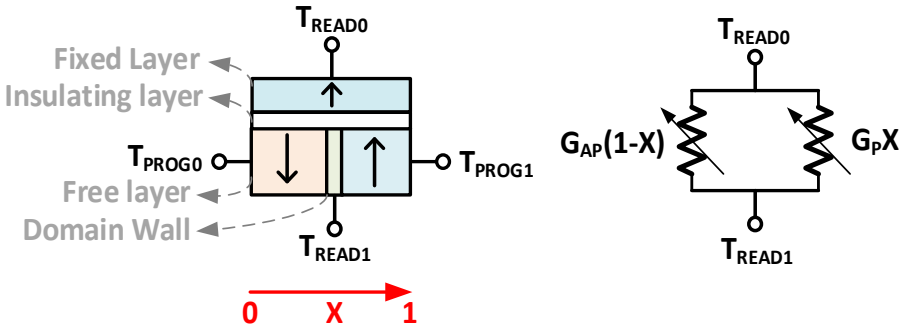


Figure 7.1 (a) Spin synapse device using horizontal charge current to program the DW position in an analog w; (b) The equivalent circuit of the conductance between the two read ports.

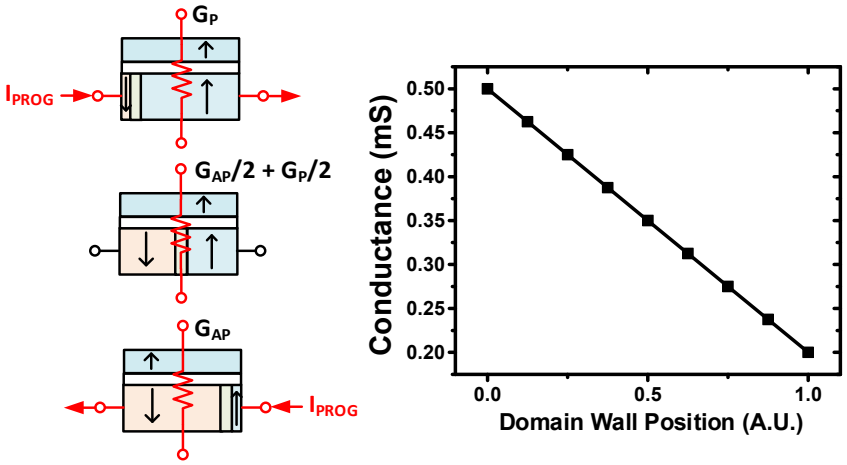
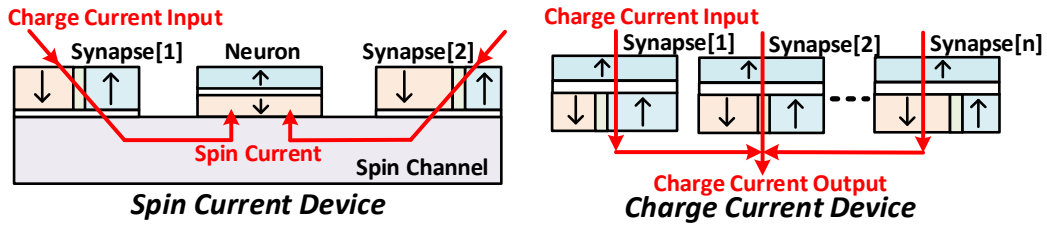


Figure 7.2 The vertical conductance of spin synapse device changes from G_P to G_{AP} according to the DW position.



(a) (b)

Figure 7.3 (a) Spin current synapse suffers from short spin diffusion distance, limiting the interconnection; (b) More charge current synapses can be connected through metal wires.

To realize analog programmability with all charge current, a spin synapse device is proposed as shown in Figure 7.1(a). The synapse device includes one fixed magnetic layer and one free layer with a DW moving inside. Due to current-induced DW motion, the DW can be moved by charge current flowing through horizontal ports T_{PROG0} and T_{PROG1}. Moreover, the DW moving distance X is linearly proportional to the current I or the time T [20], expressed as:

$$\boldsymbol{X = v * T = \frac{\beta \mu P}{\alpha e M_s}\left(\frac{I}{Area} - J_{th}\right) * T} \tag{7.1}$$

Higher current or longer time can move the DW further, which enables analog programmability of the DW position.

The vertical conductance between the two read ports T_{READ0} and T_{READ1} can be treated as the parallel connection of spin parallel conductance $G_P(1\text{-}X)$ and spin antiparallel conductance $G_{AP}X$ (Figure 7.1(b)), both of which are determined by the DW position X. Therefore, DW position can change the vertical conductance from G_P to G_{AP}, as shown in Figure 7.2(a). Because of the analog programmability of the DW position, the vertical conductance can also be varied in an analog fashion according to the DW position [87]. A Verilog-A behavioral model of the proposed synapse device is built that describes the relationship between vertical conductance and DW position as shown in Figure 7.2(b).

As shown in Figure 7.3(a), previous spin current synapses [22, 85, 86] suffer from short spin diffusion distance in the spin channel, and thus large scale connections between synapses and neurons are impossible. As no spin current is involved in the program and sensing operations of our proposed synapse device, more efficient and large scale interconnections between synapses and neurons can be realized (Figure 7.3(b)).

7.2.2 Racetrack Converter as Rectified-Linear Neuron

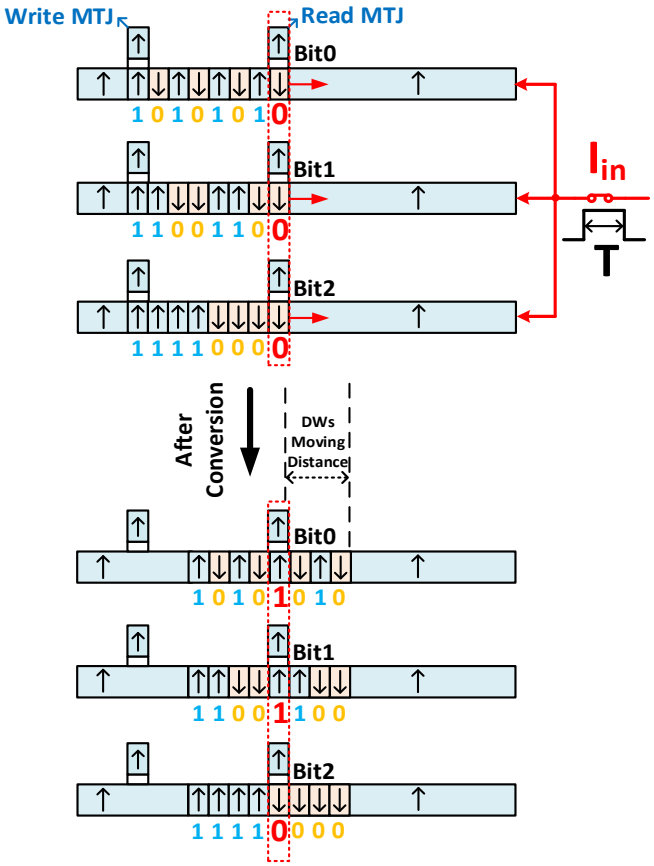


Figure 7.4 Structure of a 3b racetrack converter. Each nanowire is configured with different DWs granularity, representing an individual bit. During conversion, DWs move simultaneously and stop at a distance that is proportional to input current. Digital value is obtained by sensing the resistance of the read MTJs [23].

BTNN can use only a simple comparator neuron while an RLNN requires analog-to-digital converter (ADC) as the neuron. The large area of CMOS ADCs limits their use in RLNN. In chapter 6, a spin-based ADC was proposed with 1000× smaller area than state-of-art CMOS ADCs, while keeping similar energy efficiency. Due to its ultra-compact area and high energy efficiency, the racetrack converter is an ideal neuron for RLNN.

Figure 7.4 shows a 3-bit ADC using three racetrack magnetic nanowires. Each nanowire has one read MTJ and one write MTJ on top. Each nanowire can be configured differently with a sequence of alternating write and shift current pulses, such that each generates a single bit, from LSB to MSB. Then, the polarization of magnetic domain beneath n read MTJs represents the digitalized value from 0 to 2^n-1. During conversion, the input current can move all the DWs simultaneously. After a fixed time, the DW moving distance is determined by the input current value. Then, by sensing the resistance of the read MTJ head above each nanowire, the data can be read out as a digital value. In this way, the analog current is converted into a digital value through this racetrack converter. After read, the racetrack converter can be reset for next cycle as described in chapter 6.

7.2.3 Recurrent Neuron

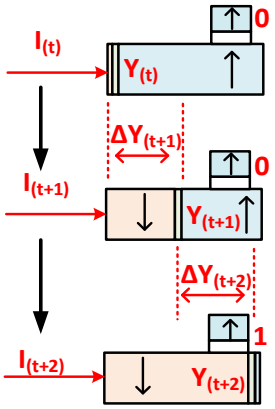


Figure 7.5 Simple recurrent DW neuron with binary-threshold output. Current-induced DW motion can store the analog DW position and perform integration for each cycle.

RNN requires analog value storage and accumulating cycle by cycle. CMOS implementations typically use a charge-storing capacitor as a recurrent neuron, however this consumes significant area and suffers from leakage concerns. DW can be a desirable analog storage element leveraging current-induced DW motion. If the injected charge current exceeds the threshold, the DW will move and the moving distance is an integration of current and time. DW motion can be stopped once current is lower than threshold and be activated again with sufficient current. Therefore, DW position can be used to memorize and accumulate analog value.

Figure 7.5 shows a simple recurrent DW neuron device. The DW starts at position $Y_{(t)}$. In the first cycle, the DW is moved by $\Delta Y_{(t+1)}$, and stops at $Y_{(t+1)}$. The device stores this DW position, and the output remains at 0 because the spin polarity of top and bottom MTJ layers are the same. But in the second cycle, the DW is moved to $Y_{(t+2)}$ from $Y_{(t+1)}$ and the output changes to 1 due to polarity flip of the bottom free layer. Therefore, the output of the neuron device is determined by not only the current input but also previous state stored in the neuron device. And thus the recurrent DW neuron can handle time-related inference tasks. This example shows a simple recurrent DW neuron device with only binary-threshold output. The racetrack converter can be modified to create a recurrent neuron with linear output functionality, because it also uses current-induced DW motion for conversion.

7.3 Neuron Network Architecture

7.3.1 Neural Network Categories

As shown in Figure 7.6, neural network function can be divided into two stages: DOT product and processing. In the first stage, input X performs DOT product with weight W stored in the synapses. In a hardware implementation, the DOT product can be realized by current summation, charge accumulation, or digital calculation.

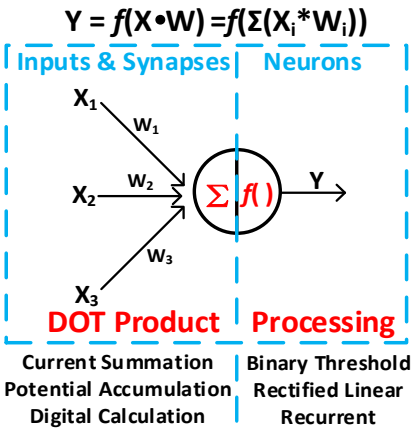


Figure 7.6 Neural network function includes DOT product and processing: inputs perform DOT product with weights stored in the synapses; neurons process the DOT product results.

After the DOT product, the neuron will perform processing on the results of the DOT product. Software implementations typically use a sigmoid or softmax function for neuron processing, however it is challenging to implement these nonlinear functions in hardware. Binary-threshold is the most common neuron function in hardware because of its simplicity: only a comparator is required (i.e, a 1-bit ADC). However, BTNN requires many more synapses and neurons to perform similar task compared to other complicated neural networks. RLNN employs an ADC as the neuron instead of a single comparator, improving its capabilities.

Both BTNN and RLNN are feed-forward neural networks which can only work on static tasks like digit recognition and image classification. To deal with time-related tasks like forecasting and phoneme recognition, conventional time-related neural networks use a shift-register at the inputs and shift the inputs cycle by cycle to add time information into the system, which are inefficient in terms of area, timing and power. Recurrent neuron itself can store the time information as internal state and generate the results based on both current inputs and previous state. Therefore, RNN can be applied to highly-efficient time-related inference.

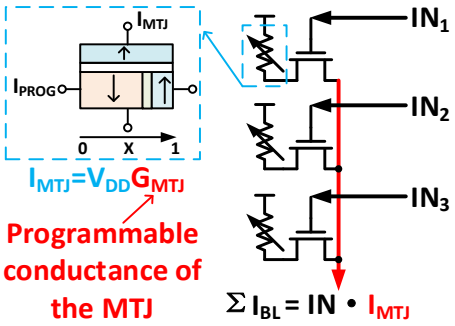


Figure 7.7 Current summation on bit-line for DOT product of inputs and weights.

7.3.2 Current Summary for DOT Product

DOT product can be implemented by current summation, charge accumulation or digital calculation (Figure 7.6). Charge accumulation suffers from leakage and digital calculation requires complicated arithmetic logic unit and large intermediate storage. Current summation is more accurate and suitable for analog computation. The proposed design using current summation for DOT product is depicted in Figure 7.7. Each spin synapse is connected to a NMOS switch, and all NMOS source terminals are shared. Input signal can turn on/off the NMOS switch: 1 connects the synapse to the bit-line; 0 isolates the synapse. If the NMOS is on, the synapse device will generate some current onto the bit-line, and all currents will be summed together on the bit-line. The current value is determined by the DW position, representing the weight. The summed current value is the DOT product of IN and I_{MTJ}. Since the NMOS on/off current ratio is more than 10^3, leakage is negligible. Moreover, current summation is a static operation, and thus it is immune to dynamic noise.

7.3.3 Cross-bar Array

Conventional spin synapses use spin current for DOT product, limiting the number of synapses connected to one neuron. The proposed spin synapse works with all charge current, and

thus multiple synapses can be connected to one bit-line, enabling massive cross-bar array configurations. Figure 7.8 shows the cross-bar synapse array structure. One spin synapse and one NMOS access device make up one cell in the array. Inputs are the word-lines of the array; weights are represented by the programmable conductance of the spin synapse device. Before any neural network operation, the weights should be programmed by horizontal current injection with varying pulse width. With low injection current, DW position can be finely controlled by considerably long pulse width.

Current summation on each bit-line performs a DOT product of inputs and weights to calculate each neuron Y_j. Neurons in a single layer are calculated in a serial fashion by adding a column MUX and a neuron DEMUX and iterating the addresses of both MUX and DEMUX. Current of a single synapse varies in the range from I_{AP} to I_P, which is barely a 2.5× range. To amplify the on/off ratio of the unit synapse current, one extra offset column is added to provide offset current I_{AP}, which increases the on/off ratio from I_P/I_{AP} to $(I_P-I_{AP})/0$. With column MUX, only one offset column is required for a full array. Also, an analog buffer is required to fix the node voltage at half VDD, such that the I_{AP} is equal on both summation and offset bit-lines. The residue current represents the DOT product results and flows into a neuron selected by neuron DEMUX for processing. The mathematic models for BTNN, RLNN and RNN are shown in the bottom of Figure 7.8.

A DW binary threshold neuron works as a current comparator. Once the residue current exceeds the threshold, the DW will move and flip the spin polarity of the bottom free layer. The proposed rectified linear neuron then uses a racetrack ADC to convert the analog residue current into a digital value and this converted digital value serves as the inputs for next layer, fully utilizing the analog residue current.

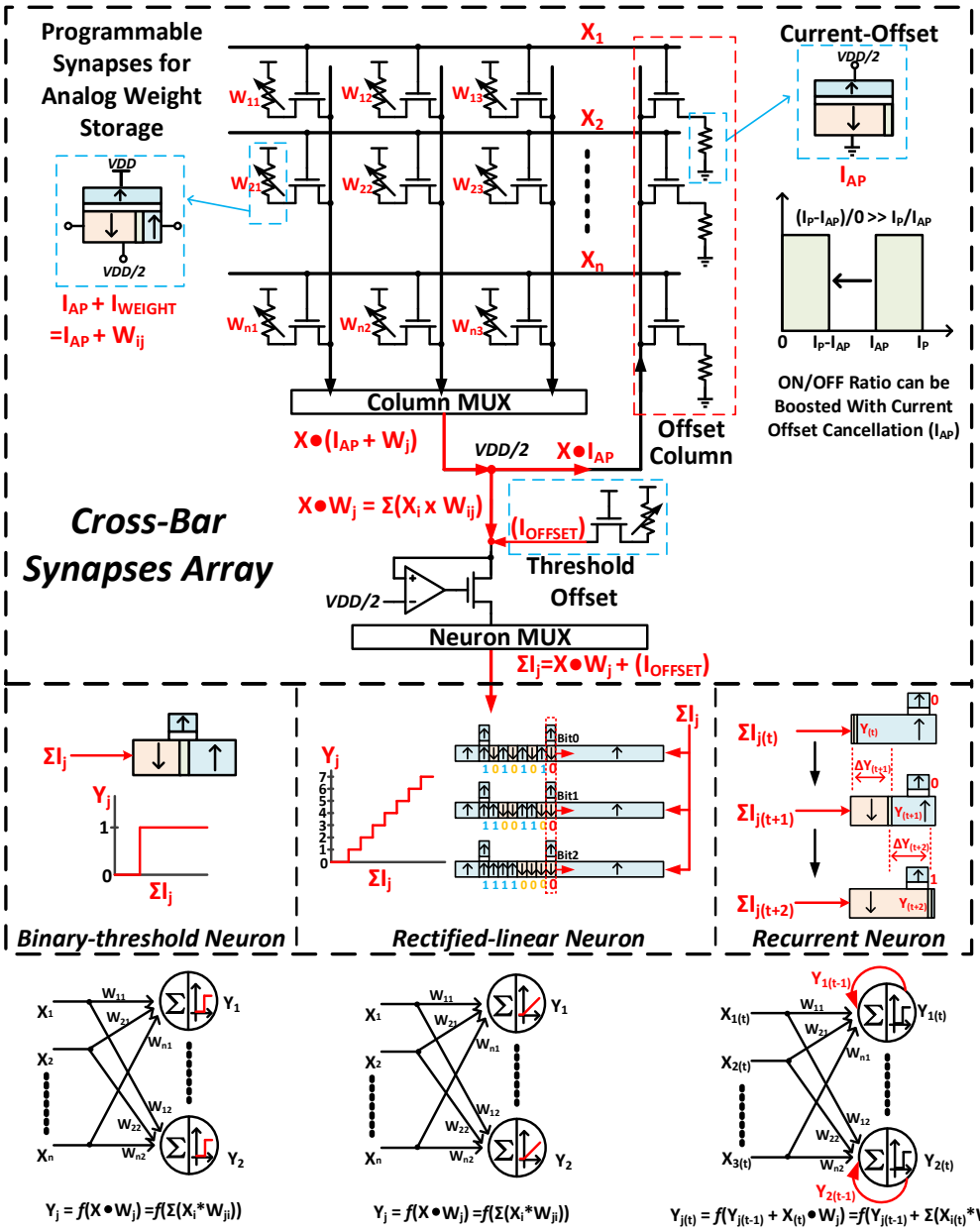


Figure 7.8 Cross-bar synapse array configuration and different neuron types. Offset column is used to improve on/off ratio. Mathematical models for BTNN, RLNN and RNN are shown in the bottom.

For the recurrent neuron, Figure 7.8 also shows a simple recurrent domain-wall neuron, storing time-related analog information as DW location. In each cycle, DW will move and stop,

and the moving distance is proportional to the residue current. In each cycle, the neuron generates a digital output. When the DW moves to the right, the neuron will be activated and output a 1. Reset can be done after activation.

As the current-induced DW motion has a current threshold, the binary-threshold neuron can use it as comparator reference. But both rectified-linear neuron and recurrent neuron require some extra offset for this threshold. This can be realized by adding an extra threshold offset device on the current subtraction node to generate a required threshold current I_{OFFSET} onto the residue current (Figure 7.8). And the DW position inside this threshold offset device should be programmed according to the current threshold of the neuron once after fabrication.

7.3.4 Majority Voting Circuit

A typical neural network system includes an input layer, several hidden layers, and an output layer as shown in Figure 7.9. For classifier applications like digit recognition, the output layer uses a majority voting circuit to make a final decision. The spin-based neural networks operate directly with current instead of voltage. Hence a majority voting circuit in this type of system must find the maximum current value among several or even tens of inputs. Conventional comparison methods need many comparators and clock cycles to find such a maximum value. Some winner-take-all circuits employ complicated analog modules. Here, we propose a simplified majority voting circuit working in time domain that more easily finds a peak value compared with using the current or voltage domain. As shown in Figure 7.9, all dedicated NMOS capacitors are discharged to ground initially. Then after signal EN goes high, the highest current path will flip the inverter first and shut off all the switches. The 0 output of the inverter represents the highest current. The operation only takes one cycle. Using just small NMOS gate capacitors the design is area-efficient and low-power.

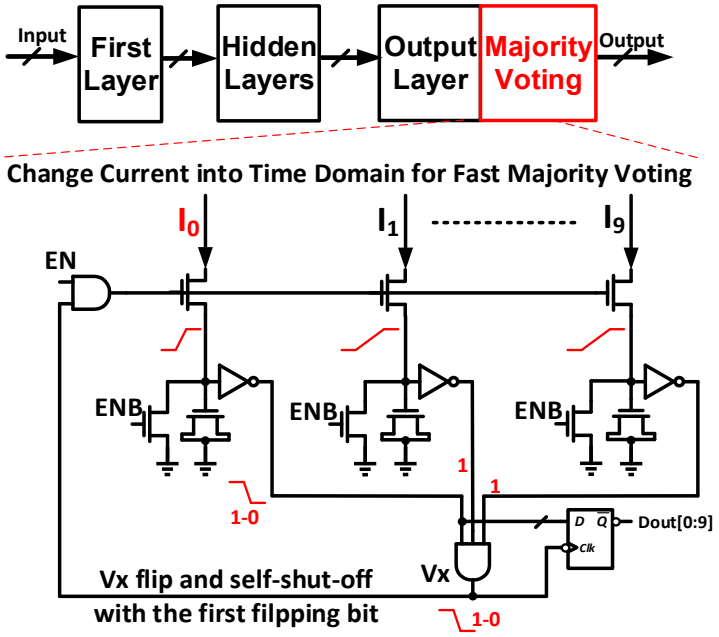


Figure 7.9 Majority voting circuit. Using cap charging instead of current comparison to find the maximum value and make final decision for recognition task.

7.3.5 Extending the RLNN to Multi-bit Input

In RLNN, input of the first hidden layer is 1-bit binary; while the following layers take 3-bit digital inputs. For 3-bit input DOT product, each bit of the input will be calculated separately as in 1-bit input case. As shown in Figure 6.10, each bit of the input fires one WL and 3-bit input takes 3 WLs. Also, 3 BLs are used for one neuron. 3 by 3 cells represent one weight: 3 of them on the diagonal have same conductance as weight; others can be programmed to provide offset. During operation, each BL will sum up the currents, then current mirrors with 1X, 2X and 4X ratios will be used to get the real total current for neurons in hidden layers or majority-voting circuit in the output layer.

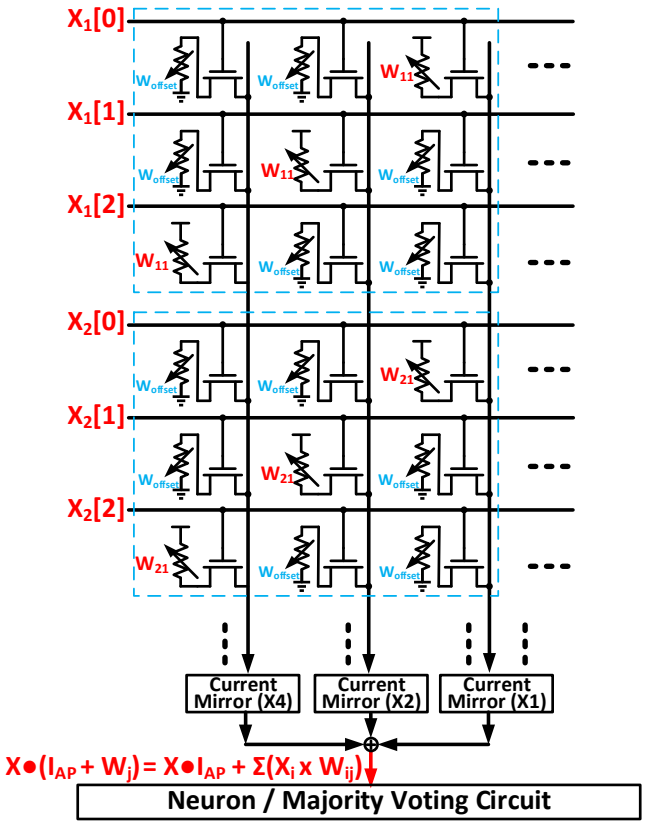


Figure 7.10 Final layer of RLNN with 3-bit digital inputs.

7.4 Simulation Results and Analysis

Compact Verilog-A models are developed for related MTJ, DW, and racetrack nanowire based on published experimental data [16-18, 20]. Co-simulation with CMOS circuits (a commercial 32nm technology) is performed in SPICE.

7.4.1 Binary-threshold and Rectified-linear Neural Networks

Both spin-based BTNN and RLNN are built containing one input layer, one hidden layer and one output layer. The neuron number in the input and output layer is tailored to serve MNIST benchmark. Figure 7.11 shows the related parameters of the synapse and neuron [20, 23, 86]. Complete 60000 MNIST digit recognition training sets are used to train both BTNN and RLNN,

and 10000 standard MNIST test sets are employed to evaluate the error rates. The evaluation test set is not included in the training set. Pseudo-gradient back-propagation algorithm [88] is employed for training. Figure 7.12 (a) and (b) show the error rate decreasing with training epochs for both BTNN and RLNN with the same number of hidden neurons (100).

Software implementations of synapse weight can use floating-point values while hardware implementations modify the floating-point values into quantized values. For BTNN, 8-bit quantized weight has little difference with floating-point weight because the binary-threshold function itself generates substantial quantization error and thus adding weight bits does not reduce error rate. However in RLNN additional weight bits serve to improve error rate.

With 100 hidden neurons, the RLNN can achieve 2.5× error rate reduction compared to BTNN (Figure 7.12(c)). Moreover, the error rate of RLNN saturates with about 5 epochs (each epoch traverses all 60000 cases and updates the weights 60000 times); while BTNN saturates much more slowly. Figure 7.12(d) compares the error rate change with synapse weight bits for both BTNN and RLNN. RLNN shows better error rate scaling with number of weight bits.

Item	Value
Technology	32nm
R_{AP}	5KΩ
R_P	2KΩ
Vdd	100mV
I_{AP}	10uA
I_P	25uA
Frequency	10MHz
Synapse Size	0.027 um^2 ($30F^2$)
Binary-threshold Neuron Size	0.4 um^2
Rectified-linear Neuron Size	2 um^2
Majority Voting Circuits Area	50 um^2
Synapse Weight Bits	3
Current Resolution	0.02uA

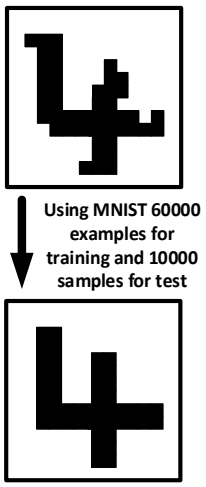


Figure 7.11 Synapse and neuron parameters used for co-simulation with CMOS; Complete MNIST digit recognition benchmark are used for training evaluation.

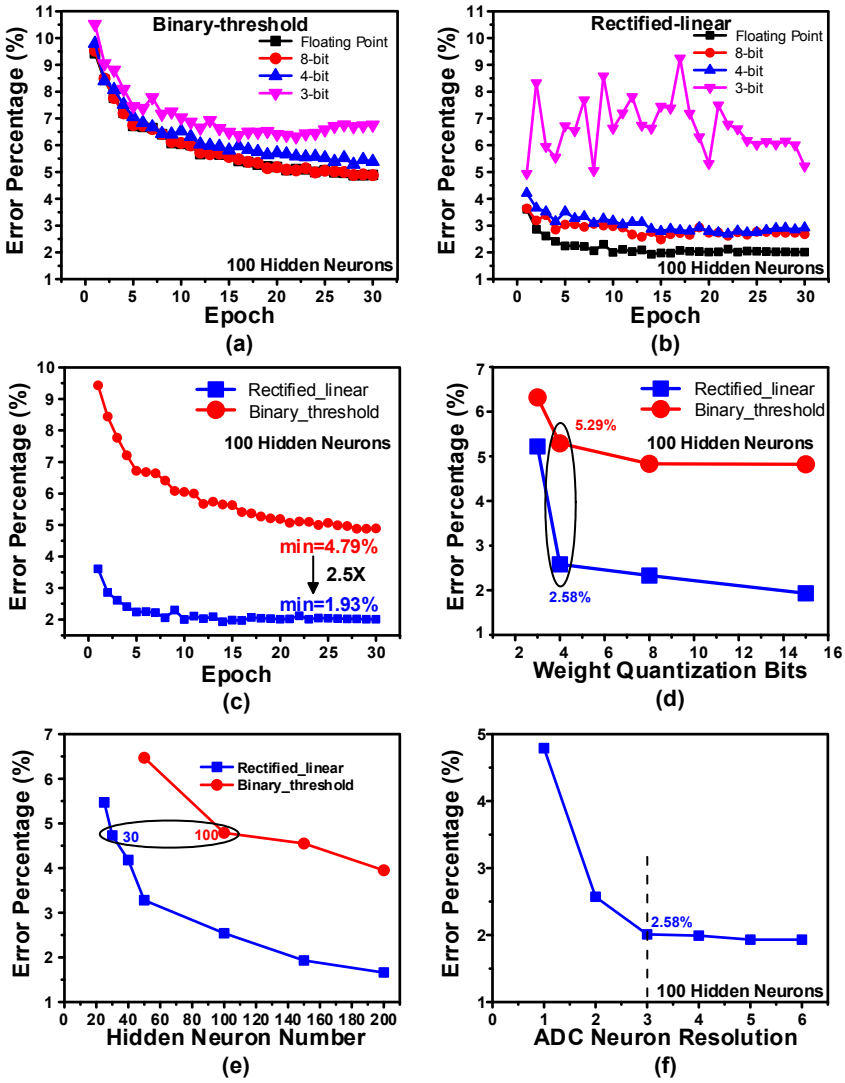


Figure 7.12. Error rate decreases with training epochs for BTNN (a) and RLNN (b). RLNN achieves 2.5× error rate reduction compared with BTNN (c). Error rate can be reduced with more weight quantization bits (d) and hidden neurons (e). Higher resolution of ADC neuron can further lower error rate of RLNN (f).

Figure 7.12(e) shows the error rate decreasing with more hidden neurons. To achieve <5% error rate, the RLNN needs only 30 hidden neurons; while BTNN requires 100 hidden neurons. Therefore, RLNN can achieve same accuracy with smaller area.

As RLNN uses an ADC as the neuron instead of a single comparator, the error rate can be reduced with more ADC resolution, as shown in Figure 7.12(f). In our experiments, a 3-bit ADC is sufficient to achieve good error rates for RLNN.

Table 7.1 compares CMOS SRAM-based BTNN, spin BTNN, and spin RLNN with the same synapse weight (4 bit). Area and energy overhead of peripheral circuits are included. While achieving the same accuracy of MNIST digit recognition task, spin-based BTNN saves ~96% area and 14% energy compared to CMOS SRAM-based BTNN. The proposed spin RLNN further reduces area by 67% and energy by 69% compared to spin BTNN.

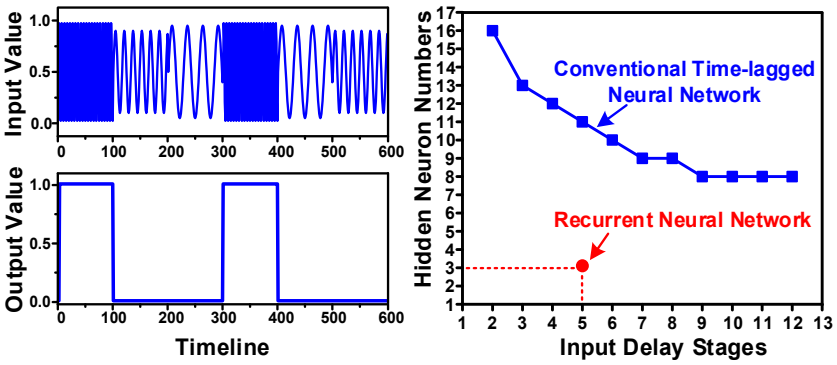
Table 7.1 Comparison between CMOS BTNN and Spin BTNN and Spin RLNN.

	CMOS BTNN	**Spin BTNN**	**Spin RLNN**
Synapse Weight	4	4	4
Hidden Neurons	100	100	30
Error Rate	4.79 %	4.79 %	4.73 %
Area	89000 um^2	3700 um^2	1220 um^2
Energy per Frame	16.6 nJ	14.2 nJ	4.4 nJ
Energy per pixel	21 pJ	18 pJ	5.6 pJ

7.4.2 Recurrent Neural Networks

RNN deals with time-related inference task. We take a simple phoneme recognition task as a training example to show the effectiveness of spin RNN. As shown in Figure 7.13(a), the input is a mixed phoneme waveform in which one specific frequency needs to be recognized, and the output of the neural network remains high once detecting this specific frequency. Conventional feed-forward binary-threshold neural network can deal with this phoneme recognition task with time-lagged inputs by adding shift registers. As shown in Figure 7.13(b), fewer hidden neurons are required as more input delay stages of the shift registers are added in a conventional time-

lagged feed-forward neural network. However, to solve the same phoneme task, RNN requires only 3 hidden neurons and 5 cycles, significantly more effective than conventional time-lagged feed-forward neural network.



(a) (b)

Figure 7.13. (a) Phoneme recognition example for training; (b) RNN requires less hidden neurons and shorter latency than conventional time-lagged neural network with shift register.

7.4.3 Discussion of Process Variation

For memory applications, 3-4 bit MTJ is challenging due to limited sensing margin and cell variation. However, in our neural network application, each BL has hundreds of cells and each neuron only need to distinguish 1-bit or 3-bit output of summed current for BTNN and RLNN, respectively. Take 100 3-bit MTJs on one BL as example, each cell generates 1-8 unit currents. For memory application, sense amplifier has to distinguish each unit of current. While in neural network, the 3-bit rectified-linear neuron only need to distinguish 100/200/…800 rather than 1/2/…8. Therefore, the sensing margin is not a problem for neural network application. Current summation of large number of cells can average out the random variation of cell current. For systematic variation, the actual measured results will show some bias, according to which the offset added to each column can be adjusted to compensate the systematic variation. Additionally,

DNN algorithm is inherently variation-tolerant. Once systematic offset of different weights is measured, it could be incorporated in the training and the training would compensate for it by adjusting the weights of the overall system.

7.5 Conclusion

A spin synapse device has been proposed with analog programmability using all charge current. The synapse devices can be placed in a cross-bar array to form a dense neural network. DOT product nano-function can be realized using current summation. With compact racetrack converter as the neuron, spin RLNN can be implemented, which saves area by 67% and energy by 69% compared to spin BTNN to solve the same MNIST digit recognition task with equal error rates. Storing the DW motion in a time-based fashion, the more complicated RNN can also be realized for time-involved inference tasks. Compared to conventional time-lagged feed-forward neural network, RNN requires fewer hidden neurons and latency cycles.

CHAPTER 8. Conclusion

8.1 Contributions of This Work

With technology scaling, memory becomes dominating part in advanced SoC in terms of area and power consumption. This thesis proposes different solutions to minimize memory area and reduce power consumption.

Chapter 2 proposes a mostly-read 5T SRAM design which achieves 7.2% area saving than 6T SRAM. With decoupled read path, read margin is improved and read access energy gets reduced. The 4Mb 5T SRAM macro is applied to a face-recognition accelerator.

Chapter 3 describes a 4+2T SRAM design that uses the N-well as a write wordline, saving 15% area than 8T SRAM. Two decoupled read paths significantly improve read noise margin, enabling reliable multi-word activation for logic operations. Using dual sense amplifiers, Boolean logic functions (AND, OR, XOR) between the two activated words can be realized. Furthermore, with separated RBL/RBLB and RWL/RWLB, the SRAM can be configured as a BCAM or TCAM, enabling searching operations.

Chapter 4 proposes a 1 Mb embedded NOR Flash memory for an ultra-low power sensor node system. Multiple low-power techniques are described to minimize the flash write power. Sub-nW voltage and current references are proposed. Also, a cross-sampling current sense amplifier is invented for sensing margin improvement. The low power NOR flash is incorporated into a complete mm-scale sensor node system to reduce sleep power and extend battery lifetime.

Chapter 5 describes a 1Mb STT-MRAM. Single-cap based offset-cancelled sense amplifier is proposed for read margin improvement and in-situ self-write termination is applied to reduce write power.

Chapter 6 proposes racetrack ADC using emerging spintronic devices based on current-induced domain wall motion. The racetrack ADC is ultra-compact, which can be applied to an ultra-high speed digital pixel sensor (DPS) imaging system.

Chapter 7 describes neural networks using these emerging spintronic devices. A spin synapse device has been developed with analog programmability using all charge current. Current summation is used to realize DOT product nano-function. With compact racetrack converter as the neuron, spin rectified-linear neural network can be implemented. Storing the DW motion in a time-based fashion, the more complicated RNN can also be realized for time-involved inference tasks.

8.2 Future Directions

There are many other possibilities to further improve memory circuit and system design based on this dissertation.

Chapter 2 and 3 propose 2 types of area-efficient low-power SRAM. Removing one of the access transistor in the 4+2T SRAM, the new 4+1T SRAM can provide more compact area than those proposed in chapter 2 and 3. It has single decoupled read port which still maintains improved read noise margin. And the read power can be saved further as data 0 won't discharge the BL which is the same as the 5T SRAM in chapter 2.

Chapter 3 also describes the concept of in-memory-computing. However, this implantation only realizes AND/OR/XOR; while other computation functions like adding and subtracting haven't been explored. Using current summation on BLs, adding and subtracting might be

realized to embed the complete ALU functions into SRAM. It will open up more possible applications with in-memory-computing.

Chapter 4 proposes an eFlash for sensor node applications. In this system, we separate the process and eFlash in different chips and stack them together. Once waked up, the process layer has to stream in all configuration data from eFlash to boot up the whole stacked system, which takes a large amount of energy and latency during the start-up period. Non-volatile processor, which embeds the non-volatile memory cells with the registers inside the processor, could significantly accelerate the system start-up and save energy consumption. Both RRAM and NRAM are promising for non-volatile processor applications due to CMOS capability, scalability, low-voltage operation and high read margin.

Chapter 5 describes a STT-MRAM design. MRAM suffers from small sensing margin. In our design, we implement the offset cancellation inside the latch-based second-stage sense amplifier. However, the first stage (PMOS headers) still contributes variation during read operation. To mitigate the variation of the PMOS headers, cross-coupled current-sampling, which has been proposed in chapter 4, can be applied to alleviate the variation of the PMOS headers, further improving sensing margin.

Chapter 6 and 7 explore the possibilities of applying emerging spintronic devices to analog computing and neuromorphic computing. Other applications such as stochastic computing and in-memory-computing, can also benefit from the emerging spintronic devices due to their compact area, non-volatility and great endurance.

8.3 Related Publications

[1] **Qing Dong**, David Blaauw, and Dennis Sylvester, "A 1.02nW PMOS-Only, Trim-Free Current Reference with 282ppm/°C from -40°C to 120°C and 1.6% within-Wafer Inaccuracy," IEEE European Solid-State Circuits Conference (ESSCIRC), 2017

[2] **Qing Dong**, Supreet Jeloka, Mehdi Saligane, Yejoong Kim, Masaru Kawaminami, Akihiko Harada, Satoru Miyoshi, David Blaauw, and Dennis Sylvester, "A 0.3V VDDmin 4+2T SRAM for Searching and In-Memory Computing Using 55nm DDC Technology," IEEE Symposium on VLSI Circuits (VLSIC), 2017

[3] **Qing Dong**, Yejoong Kim, Inhee Lee, Myungjoon Choi, Ziyun Li, Jingcheng Wang, Kaiyuan Yang, Yen-Po Chen, Gyouho Kim, Junjie Dong, Minchang Cho, Yun-Sheng Chen, Yu-Der Chih, David Blaauw, and Dennis Sylvester, "A 1Mb Embedded NOR Flash Memory with 39uW Program Power for mm-Scale High-Temperature Sensor Nodes," IEEE International Solid-State Circuits Conference (ISSCC), 2017

[4] **Qing Dong**, Kaiyuan Yang, Laura Fick, David Blaauw, Dennis Sylvester, "Binary-threshold, Rectified-linear and Recurrent Neural Networks Built with Spintronic Devices," IEEE International Symposium on Circuits and System (ISCAS), 2017

[5] ***Qing Dong**, *Dongsuk Jeon, Yejoong Kim, Xiaolong Wang, Shuai Chen, Hao Yu, David Blaauw, Dennis Sylvester, "A 23mW Face Recognition Processor with Mostly-Read 5T Memory in 40nm CMOS," IEEE Journal of Solid-State Circuits, (JSSC), 2017, (*Equally Contributed)

[6] **Qing Dong**, Kaiyuan Yang, Laura Fick, David Fick, David Blaauw, Dennis Sylvester, "A Low Power and Compact Data Converter Using Racetrack Spintronic Devices for High-Speed Imaging System," IEEE Transactions on VLSI Systems (TVLSI), 2017

[7] **Qing Dong**, Kaiyuan Yang, David Blaauw and Dennis Sylvester, "A 114-pW PMOS-Only, Trim-Free Voltage Reference with 0.26% within-Wafer Inaccuracy for nW Systems," IEEE Symposium on VLSI Circuits (VLSIC), 2016

[8] Dongsuk Jeon, **Qing Dong**, Yejoong Kim, Xiaolong Wang, Shuai Chen, Hao Yu, David Blaauw, Dennis Sylvester,"A 23mW Face Recognition Accelerator in 40nm CMOS with Mostly-Read 5T Memory," IEEE Symposium on VLSI Circuits (VLSIC), 2015

[9] **Qing Dong**, Kaiyuan Yang, Laura Fick, David Fick, David Blaauw, Dennis Sylvester, "Racetrack Converter: A Low Power and Compact Data Converter Using Racetrack Spintronic Devices," IEEE International Symposium on Circuits and System (ISCAS), 2015

BIBLIOGRAPHY

[1] http://isscc.org/doc/2017/ISSCC2017_TechTrends.pdf

[2] J. Chang, et al., “The 65-nm 16-MB Shared On-Die L3 Cache for the Dual-Core Intel Xeon Processor 7100 Series”, IEEE Journal of Solid-State Circuits, vol. 43, no. 4, pp. 92-94, Apr. 2007.

[3] M. Khellah, et al., “Wordline & Bitline Pulsing Schemes for Improving SRAM Cell Stability in Low-Vcc 65nm CMOS Designs,” IEEE Symposium on VLSI Circuits 2006, pp. 9-10.

[4] K. Takeda, et al., “Multi-Step Word-Line Control Technology in Hierarchical Cell Architecture for Scaled-Down High-Density SRAMs,” IEEE Journal of Solid-State Circuits, vol. 46, no. 4, pp. 806-814, Apr. 2011.

[5] M. Bhargava, et al., “Low VMIN 20nm Embedded SRAM with Multi-voltage Wordline Control based Read and Write Assist Techniques,” IEEE Symposium on VLSI Circuits 2014, pp. 1-2.

[6] M.-F. Chang, et al., “A Sub-0.3 V Area-Efficient L-Shaped 7T SRAM With Read Bitline Swing Expansion Schemes Based on Boosted Read-Bitline, Asymmetric-Vth Read-Port, and Offset Cell VDD Biasing Techniques,” IEEE Journal of Solid-State Circuits, vol. 48, no. 10, pp. 2558-2569, Oct. 2013.

[7] T. Suzuki, et al., “A Stable 2-Port SRAM Cell Design Against Simultaneously Read/Write-Disturbed Accesses,” IEEE Journal of Solid-State Circuits, vol. 43, no. 9, pp. 2109-2119, Sep. 2008.

[8] N. Verma, and A. P. Chandrakasan, “A 256 kb 65 nm 8T Subthreshold SRAM Employing Sense-Amplifier Redundancy,” IEEE Journal of Solid-State Circuits, vol. 43, no. 1, pp. 141-149, Jan. 2008.

[9] L. Chang, et al., “An 8T-SRAM for Variability Tolerance and Low-Voltage Operation in High-Performance Caches,” IEEE Journal of Solid-State Circuits, vol. 43, no. 4, pp. 956-963, Apr. 2008.

[10] B. H. Calhoun, and A. Chandrakasan, “A 256kb Sub-threshold SRAM in 65nm CMOS,” IEEE International Solid-State Circuits Conference 2006, pp. 628-629.

[11] I. Chang, J. Kim, S. Park, and K. Roy, “A 32kb 10T Subthreshold SRAM Array with Bitinterleaving and Differential Read Scheme in 90nm CMOS,” IEEE International Solid-State Circuits Conference, Feb. 2008, pp. 388-389.

[12] S. Nalam, and B. H. Calhoun, “Asymmetric Sizing in a 45nm 5T SRAM to Improve Read Stability over 6T,” IEEE Custom Integrated Circuits Conference 2008, pp. 709-712.

[13] S. Jeloka, N. B. Akesh, D. Sylvester and D. Blaauw, "A 28 nm Configurable Memory (TCAM/BCAM/SRAM) Using Push-Rule 6T Bit Cell Enabling Logic-in-Memory," IEEE Journal of Solid-State Circuits, vol. 51, no. 4, pp. 1009-1021, April 2016.

[14] Y. Lee, et al., “A Modular 1mm3 Die-Stacked Sensing Platform with Optical Communication and Multi-Modal Energy Harvesting,” IEEE International Solid-State Circuits Conference 2012, pp. 229-243.

[15] G. Kim, et al., “A Millimeter-Scale Wireless Imaging System with Continuous Motion Detection and Energy Harvesting,” IEEE Symposium on VLSI Circuits 2014.

[16] D. Chiba, et al., “Control of Multiple Magnetic Domain Walls by Current in a Co/Ni Nano-Wire,” Applied Physics Express, 2010.

[17] S. Fukami, et al., “High-speed and reliable domain wall motion device: Material design for embedded memory and logic application,” IEEE Symposium on VLSI Technology, 2012.

[18] L. Thomas, et al., “Racetrack Memory: a high-performance, low-cost, non-volatile memory based on magnetic domain walls,” IEEE International Electron Devices Meeting, 2011.

[19] A. J. Annunziata, et al., “Racetrack Memory Cell Array with Integrated Magnetic Tunnel Junction Readout,” IEEE International Electron Devices Meeting, 2011.

[20] Y. Zhang, et al., “Perpendicular-magnetic-anisotropy CoFeB racetrack memory,” Journal of Applied Physics, 2012.

[21] M. Sharad, et al., “Spin-Based Neuron Model with Domain-Wall Magnets as Synapse,” Trans. on Nanotechnology, 2012, pp. 843-853.

[22] M. Sharad, et al., “Ultra Low Power Associative Computing with Spin Neurons and Resistive Crossbar Memory,” ACM/IEEE Design Automation Conference, 2013.

[23] Q. Dong, et al., “Racetrack Converter: A Low Power and Compact Data Converter Using Racetrack Spintronic Devices,” IEEE International Symposium on Circuits and Systems, 2015.

[24] H. Noguchi, et al., “A 250-MHz 256b-I/O 1-Mb STT-MRAM with Advanced Perpendicular MTJ based Dual Cell for Nonvolatile Magnetic Caches to Reduce Active Power of Processors,” IEEE Symposium on VLSI Circuits, 2013.

[25] D. Kim, et al., “A 1.85fW/bit Ultra Low Leakage 10T SRAM with Speed Compensation Scheme,” IEEE International Symposium on Circuits and Systems, 2011.

[26] T. Kim, J. Liu, and C. H. Kim, “A Voltage Scalable 0.26 V, 64 kb 8T SRAM With Vmin Lowering Techniques and Deep Sleep Mode,” IEEE Journal of Solid-State Circuits, vol. 44, no. 6, pp. 1785-1795, Jun. 2009.

[27] A. J. Bhavnagarwala, et al., “A Sub-600-mV, Fluctuation Tolerant 65-nm CMOS SRAM Array With Dynamic Cell Biasing,” IEEE Journal of Solid-State Circuits, vol. 43, no. 4, pp. 946-955, Apr. 2008.

[28] E. Karl, et al., “A 4.6 GHz 162 Mb SRAM Design in 22 nm Tri-Gate CMOS Technology With Integrated Read and Write Assist Circuitry,” IEEE Journal of Solid-State Circuits, vol. 48, no. 1, pp. 150-158, Jan. 2013.

[29] D. Kim and M. Seok, “Fully Integrated Low-Drop-Out Regulator Based on Event-Driven PI Control,” IEEE International Solid-State Circuits Conference 2016, pp. 148-149.

[30] M. Kang, et al., “In-Memory Computing Architectures for Sparse Distributed Memory,” Transaction on Biomedical Circuits and Systems, vol. 10, no.4, pp. 855-863, 2016.

[31] D. Jeon, et al., “A 23mW Face Recognition Accelerator in 40nm CMOS with Mostly-Read 5T Memory,” IEEE Symposium on VLSI Circuits 2015.

[32] A. Agarwal et al., “A 128×128b high-speed wide-and match-line content addressable memory in 32 nm CMOS,” IEEE European Solid-State Circuits Conference 2011, pp. 83-86.

[33] M.-F. Chang et al., “A Process Variation Tolerant Embedded Split-Gate Flash Memory Using Pre-Stable Current Sensing Scheme,” IEEE Journal of Solid-State Circuits, vol. 44, no.3, pp. 855-863, Mar. 2009.

[34] N. Derhacobian et al., “Power and Energy Perspectives of Nonvolatile Memory Technologies,” Proceedings of the IEEE, 2010. vol. 98, no.2, pp. 283-298, Feb. 2010.

[35] Q. Dong et al., “A 114-pW PMOS-Only, Trim-Free Voltage Reference with 0.26% within-Wafer Inaccuracy for nW Systems,” IEEE Symposium on VLSI Circuits 2016, pp. 98-99.

[36] Y. Kim, Yoonmyung Lee, Dennis Sylvester, David Blaauw, “SLC: Split-Control Level Converter for Dense and Stable Wide-Range Voltage Conversion,” IEEE European Solid-State Circuits Conference 2012.

[37] G. Ge et al., “A Single-Trim CMOS Bandgap Reference With a 3σ Inaccuracy of 0.15% from 40˚C to 125˚C,” IEEE Journal of Solid-State Circuits, vol. 46, no. 11, pp. 2693-2701, Nov. 2011.

[38] M. Seok, et al., “A Portable 2-Transistor Picowatt Temperature-Compensated Voltage Reference Operating at 0.5 V,” IEEE Journal of Solid-State Circuits, vol. 47, no. 10, pp. 2534-2545, Oct. 2012.

[39] I. Lee, et al., “Low Power Battery Supervisory Circuit with Adaptive Battery Health Monitor,” IEEE Symposium on VLSI Circuits 2014.

[40] A. Shrivastava, et al., “A 32nW Bandgap Reference Voltage Operational from 0.5V Supply for Ultra-Low Power Systems,” IEEE International Solid-State Circuits Conference 2015, pp. 94-96.

[41] Y. Osaki, et al., “1.2-V Supply, 100-nW, 1.09-V Bandgap and 0.7-V Supply, 52.5-nW, 0.55-V Subbandgap Reference Circuits for Nanowatt CMOS LSIs,” IEEE Journal of Solid-State Circuits, vol. 48, no. 6, pp. 1530-1538, 2013.

[42] V. Ivanov, et al, “An Ultra-Low Power Bandgap Operational at Supply From 0.75 V”, IEEE Journal of Solid-State Circuits, vol. 47, no. 7, pp. 1515-1523, 2012.

[43] T. Hirose, et al., “A Nano-Ampere Current Reference Circuit and its Temperature Dependence Control by using Temperature Characteristics of Carrier Mobilities,” IEEE European Solid-State Circuits Conference 2010.

[44] M. Choi, et al., “A 23pW, 780ppm/°C Resistor-less Current Reference Using Subthreshold MOSFETs,” IEEE European Solid-State Circuits Conference 2014.

[45] S. Chouhan, et al., “A 0.67-µW 177-ppm/◦C All-MOS Current Reference Circuit in a 0.18-µm CMOS Technology, IEEE Transaction on Circuits and Systems II, vol. 63, no. 8, pp. 723-727, Aug. 2016.

[46] H. Kayahan, et al., “Wide Range, Process and Temperature Compensated Voltage Controlled Current Source,” IEEE Transaction on Circuits and Systems I, vol.60, no.5, pp. 1345-1353, May 2013.

[47] J. Lee, et al, “A 1.4-µW 24.9-ppm/°C Current Reference With Process-Insensitive Temperature Compensation in 0.18-µm CMOS,” IEEE Journal of Solid-State Circuits, vol. 47, no. 10, pp. 2527-2533, 2012.

[48] E. Mauricio, et al., “A 2-nW 1.1-V Self-Biased Current Reference in CMOS Technology,” IEEE Transaction on Circuits and Systems II, vol. 52, no. 2, pp. 61-65, Feb. 2005.

[49] H. Mitani et al., “A 90nm Embedded 1T-MONOS Flash Macro for Automotive Applications with 0.07mJ/8kB Rewrite Energy and Endurance Over 100M Cycles Under Tj of 175°C,” ,” IEEE International Solid-State Circuits Conference 2016, pp. 140-142.

[50] K. C. Chun, et al., “A Scaling Roadmap and Performance Evaluation of In-Plane and Perpendicular MTJ Based STT-MRAMs for High-Density Cache Memory,” IEEE Journal of Solid-State Circuits, vol. 48, no. 2, pp. 598-610, Feb. 2013.

[51] B. Giridhar, Nathan Pinckney, Dennis Sylvester, David Blaauw, “A Reconfigurable Sense Amplifier with Auto-Zero Calibration and Pre-Amplification in 28nm CMOS,” IEEE International Solid-State Circuits Conference 2014.

[52] Q. Dong, et al., “A 1Mb Embedded NOR Flash Memory with 39µW Program Power for mm-Scale High-Temperature Sensor Nodes”, IEEE International Solid-State Circuits Conference 2017.

[53] J. Javanifard, et al., “A 45nm Self-Aligned-Contact Process 1Gb NOR Flash with 5MB/s Program Speed,” IEEE International Solid-State Circuits Conference 2008.

[54] H. Noguchi, et al., “A 3.3ns-Access-Time 71.2μW/MHz 1Mb Embedded STT-MRAM Using Physically Eliminated Read-Disturb Scheme and Normally-Off Memory Architecture,” IEEE International Solid-State Circuits Conference 2015.

[55] H.-C. Yu, et al., “Cycling Endurance Optimization Scheme for 1Mb STT-MRAM in 40nm Technology,” IEEE International Solid-State Circuits Conference 2013.

[56] T. Ohsawa, et al., “1Mb 4T-2MTJ Nonvolatile STT-RAM for Embedded Memories Using 32b Fine-Grained Power Gating Technique with 1.0ns/200ps Wake-up/Power-off Times,” ,” IEEE Symposium on VLSI Circuits 2012.

[57] K. Tsuchida, et al., “A 64Mb MRAM with Clamped-Reference and Adequate-Reference Schemes,” IEEE International Solid-State Circuits Conference 2010.

[58] R. Nebashi, et al., “A 90nm 12ns 32Mb 2T1MTJ MRAM,” IEEE International Solid-State Circuits Conference 2009.

[59] S. Kleinfelder, et al., “A 10000 Frames/s CMOS Digital Pixel Sensor,” IEEE Journal of Solid-State Circuits, vol. 36, no. 12, pp. 2049-2059, 2001.

[60] X. Li, et al., “Adaptive Post-Silicon Tuning for Analog Circuits: Concept, Analysis and Optimization,” IEEE/ACM International Conference on Computer-Aided Design 2007.

[61] M. Sharad, et al., “Low Power and Compact Mixed-Mode Signal Processing Hardware using Spin-Neurons,” IEEE International Symposium on Quality Electronic Design 2013.

[62] M. Furuta, et al., “A High-Speed, High-Sensitivity Digital CMOS Image Sensor with a Global Shutter and 12-bit Column-Parallel Cyclic A/D Converters,” IEEE Journal of Solid-State Circuits, vol. 42, no. 4, pp. 766-774, 2007.

[63] S. Lim, et al., “A High-Speed CMOS Image Sensor with Column-Parallel Two-Step Single-Slope ADCs,” IEEE Transaction on Electron Devices, vol. 56, no. 3, pp. 393-398, 2009.

[64] N. Ben-Romdhane, “Design and Analysis of Racetrack Memory Based on Magnetic Domain Wall Motion in Nanowires,” IEEE/ACM International Symposium on Nanoscale Architectures, 2014.

[65] S. Fukami, et al., “Relation between critical current of domain wall motion and wire dimension in perpendicularly magnetized Co/Ni nanowires,” Applied Physics Letters, vol. 95, no. 23, 2009.

[66] A. Yamaguchi, et al., “Reduction of Threshold Current Density for Current-Driven Domain Wall Motion using Shape Control,” Japanese Journal of Applied Physics, vol. 45, no. 5A, pp. 3850-3853, 2006.

[67] S. Fukami, et al., “Low-Current Perpendicular Domain Wall Motion Cell for Scalable High-Speed MRAM,” IEEE Symposium on VLSI Technology, 2009.

[68] C. Augustine, et al., “Numerical Analysis of Domain Wall Propagation for Dense Memory Arrays,” IEEE International Electron Devices Meeting 2011.

[69] S. Fukami, et al., “Domain Wall Motion Device for Nonvolatile Memory and Logic — Size Dependence of Device Properties,” IEEE Transaction on Magnetics, vol. 50, no. 11, Nov. 2014

[70] D. Morris, et al., “mLogic: Ultra-Low Voltage Non-Volatile Logic Circuits Using STT-MTJ Devices,” ACM/IEEE Design Automation Conference, 2012.

[71] B. B. Aein, et al., “Proposal for an all-spin logic device with built-in memory,” Nature Nanotechnology, vol. 5, pp. 266-270, Apr. 2010.

[72] Y. Zhang, et al., “STT-RAM Cell Design Optimization for Persistent and Non-Persistent Error Rate Reduction: A Statistical Design View,” IEEE IEEE/ACM International Conference on Computer-Aided Design 2011.

[73] W. C. Black, et al., “Time Interleaved Converter Arrays,” IEEE Journal of Solid-State Circuits, vol. 15, no. 6, pp. 1022-1029, 1980.

[74] J. M. Musicer, et al., “MOS Current Mode Logic for Low Power, Low Noise CORDIC Computation in Mixed-signal Environments,” IEEE/ACM International Symposium on Low Power Electronics and Design 2000.

[75] M. J. Hall, et al., “Noise Analysis of a Current-Mode Read Circuit for Sensing Magnetic Tunnel Junction Resistance,” IEEE International Symposium on Circuits and Systems, 2011.

[76] R. Nebashi, et al., “A Content Addressable Memory Using Magnetic Domain Wall Motion Cells,” IEEE Symposium on VLSI Circuits, 2011.

[77] T. Suzuki, et al., “Low-Current Domain Wall Motion MRAM with Perpendicularly Magnetized CoFeB/MgO Magnetic Tunnel Junction and Underlying Hard Magnets,” IEEE Symposium on VLSI Technology, 2013.

[78] S. Fukami, et al., “Scalability Prospect of Three-Terminal Magnetic Domain-Wall Motion Device,” IEEE Transaction on Magnetics, vol. 48, no. 7, pp. 2152-2157, Jul. 2012

[79] S. Fukami, et al., “20-nm magnetic domain wall motion memory with ultralow-power operation,” IEEE IEEE International Electron Devices Meeting 2013.

[80] S. Motaman, et al., “Adaptive Write and Shift Current Modulation for Process Variation Tolerance in Domain Wall Caches,” IEEE Transaction on VLSI, vol. 24, no. 3, pp. 944-953, Mar 2016.

[81] R. Dorrance, et al., “Scalability and Design-Space Analysis of a 1T-1MTJ Memory Cell for STT-RAMs,” IEEE Transaction on Electron Devices, vol. 59, no. 4, pp. 878-887, Apr. 2012.

[82] J. A. Fredenburg, et al., “Statistical Analysis of ENOB and Yield in Binary Weighted ADCs and DACS with Random Element Mismatch,” IEEE Transaction on Circuits and Systems I, vol. 59, no. 7, pp. 1396-1408, July 2012

[83] P. Harpe, et al., “A 12fJ/Conversion-Step 8bit 10MS/s Asynchronous SAR ADC for Low Energy Radios,” IEEE European Solid-State Circuits Conference 2010.

[84] K. Yoshioka, et al., “An 8bit 0.35-0.8V 0.5-30MS/s 2bit/step SAR ADC with Wide Range Threshold Configuring Comparator,” IEEE European Solid-State Circuits Conference 2012.

[85] M. Sharad, et al., “Spin Neuron for Ultra Low Power Computational Hardware,” Device Research Conference, 2012.

[86] M. Sharad, et al., “Spin-Based Neuron Model with Domain-Wall Magnets as Synapse,” IEEE Transaction on Nanotechnology, vol. 11, no. 4, pp. 843-853, 2012.

[87] X. Wang, et al., “Spintronic Memristor through Spin-Torque-Induced Magnetization Motion,” IEEE Electron Device Letters, vol. 30, no. 3, pp. 294-297, 2009.

[88] R. M. Goodman, et al., “A Learning Algorithm for Multi-Layer Perceptrons with Hard-Limiting Threshold Units,” Proceedings of IEEE Workshop on Neural Networks for Signal Processing, 1994.