Study on Energy Minimization of Intermittent Operation Applications using Non-Volatile Power Gating

February 2024

Aika Kamei

Study on Energy Minimization of Intermittent Operation Applications using Non-Volatile Power Gating

Aika Kamei

Keio University

A thesis for the degree of Ph.D. in Engineering
Under the supervision of **Prof. Hideharu Amano**

Graduate School of Science and Technology
Keio University

February 2024

Abstract

The surge in computing necessitates advancements in low-power LSIs for edge computing. For battery-powered edge devices, low-power LSIs are crucial, as energy consumption directly impacts recharge frequency, device lifespan, and maintenance costs. To design energy-efficient chips, it's essential to select technologies that are best suited to the operating characteristics of the target application, and this requires energy consumption estimation during the design phase. In intermittent operation applications commonly seen in edge computing, which periodically alternate between active and idle states, power gating is effective in reducing leak power during inactivity. Furthermore, when data retention is necessary, nonvolatile power gating (NVPG) using nonvolatile memory proves beneficial.

This study aims to construct an energy model based on actual measurement data, using the most promising emerging nonvolatile memory technology today, Spin-Transfer Torque Magnetic Tunnel Junctions (STT-MTJs), implemented in a chip fabricated with a 40-nm MTJ/CMOS hybrid process for estimating energy consumption in intermittent operation applications within NVPG scenarios. The challenge with STT-MTJ involves the variability in the MTJ's switching characteristics, which leads to a significant energy to switching the MTJ states, called "store energy." The proposed model incorporates and models this variability assuming a normal distribution based on observations of the implemented chips, enabling sufficiently practical accuracy in energy estimation. Additionally, this energy model can be applied to build energy models for alternative options such as conventional volatile FFs, retention FFs using high-V_{th} MOS transistors, and other NVFFs, allowing for a quantitative comparison of these various FF technologies.

Moreover, this study proposes a workflow for breakeven analysis using the defined energy model, considering the energy reduction effects and overheads of various FFs, to select the optimal technology for energy minimization. This provides a quantitative basis for decision-making in selecting the most suitable NVFF and its best operational method according to the target intermittent operation application. This encourages designers to more actively adopt NVPG using MTJ-based NVFFs.

Contents

List of Figures

List of Tables

List of important abbreviations

BUP bit update probability

CGRA coarse-grained reconfigurable arrays/architecture

DAS data aware store

FF flip-flop

IoT internet of things

MTJ magnetic tunnel junction

NOF normally-off computing

NVFF nonvolatile flip-flop

NVM nonvolatile memory

NVPG nonvolatile power gating

OSS one-step store

PD power domain

PG power gating

PR pass rate

PSM-NVFF pseudo-spin-MOSFET nonvolatile flip-flop

RFF retention flip-flop

SD store domain

SSR-NVFF Split Store/Restore nonvolatile flip-flop

STT spin-transfer torque

TSS two-step store

VR-NVFF verify-and-retryable nonvolatile flip-flop

1

Introduction

Over the past two decades, despite the end of Dennard scaling and the slowdown of Moore's Law, high-performance computing has continued to evolve at a relatively constant rate of improvement due to ongoing innovation, doubling every 1.2 years, while the rate of performance-per-watt advancements in the field of computing has decreased to less than half that of the rate of performance increase. If we continue to advance at the historical rate, it is possible to achieve a zettaflop of computing in approximately ten years. However, generating half a gigawatt of electricity, which is equivalent to the output of half a nuclear power plant, is obviously not a feasible option.

To meet the continually increasing demands for computing, one viable solution is edge computing, or the concept of the Internet of Things (IoT). This approach involves processing data close to its source, either on data collection devices or on computers located near these devices, instead of traditional data centers or the cloud. Edge computing optimizes computing efficiency by reducing data movement and associated processing, since it performs the necessary computations closer to where they are needed. Furthermore, by offloading computation from cloud data centers to network edges and edge nodes, edge computing offers significant advantages in terms of lower latency and enhanced security [10].

1.1 The call for energy-efficient IoT devices

The energy consumption of edge devices in the edge computing environment is enormous and just as crucial as that of cloud data centers, considering the billions of devices deployed. Many edge devices, often constrained by battery power, require energy-aware edge computing to extend their lifespan, ensure service quality, and enhance system performance, all within a specific power budget. Consequently, the importance of research focused on improving energy efficiency in edge environments has been escalating in recent years [10,11]. Research in energy efficiency in edge computing encompasses a range of interconnected research challenges and directions. These include architecture, operating systems, middleware, application services, and computation offloading. Numerous studies have been conducted to improve the energy efficiency of computations by introducing new computational paradigms such as approximate computing and computing-in-memory [11–14].

On the other hand, with the advancement of semiconductor process technology and the miniaturization of transistors, the leakage current that occurs during idle periods has become a significant factor in energy loss. Power gating, a technique that reduces leakage current by cutting off the power supply to inactive logic blocks and memory, has been proposed. The utilization of nonvolatile memory (NVM), which retains data even without power, can sustain memory content while no leakage current is wasted. By storing data in a nonvolatile element encapsulated in memory, it is possible to reduce overheads such as data evacuation and recovering associated with power gating.

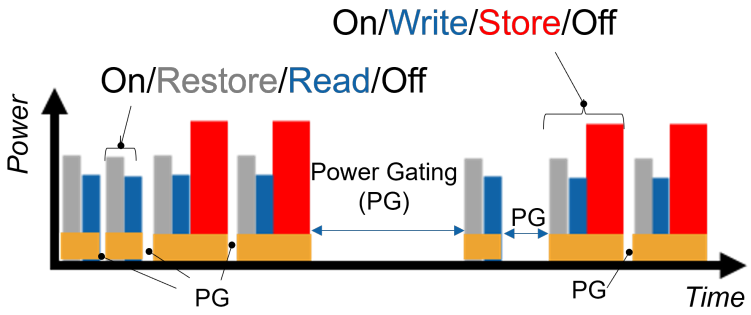
1.2 Nonvolatile memory in edge computing

Memory can be classified into volatile memory, which loses data when power is turned off, and NVM, which can retain data even without power supply. Ideally, if all the memory elements in a computer system could be made nonvolatile, it would be possible to completely eliminate wasteful leakage current during idle periods of applications. However, the current situation is as follows.

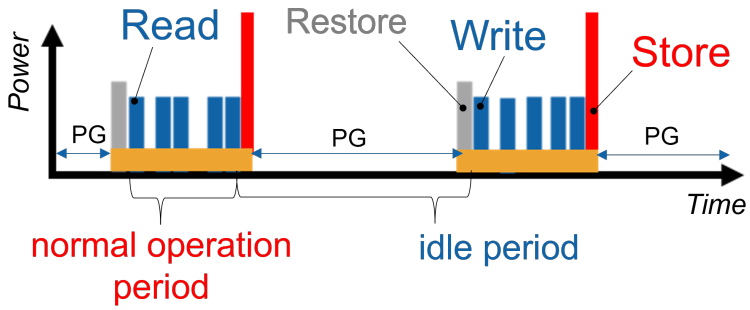
Presently, NVM used in edge devices, such as EEPROM (electrically erasable programmable read-only memory) and eFlash (embedded Flash memory), primarily stores application programs and operates mainly as read-only-memory during execution. Due to the limitations of further miniaturization (below 28nm or 22nm), eFlash is beginning to be replaced by emerging NVM technology, such as eMRAM (embedded magnetoresistive random-access memory) [15]. Most eMRAM employs spintronics-based nonvolatile elements STT-MTJ (spin-transfer torque magnetic tunnel junction), and STT-MTJ is one of the most industrially mature technologies among other emerging nonvolatile

technologies like ReRAM (resistive RAM), FeRAM (ferroelectric RAM), and PCRAM (phase-change RAM).

However, faster and more frequently accessed data memories and registers, such as SRAM (static RAM), FFs (flip-flops), and Latches remain volatile. This is because there is currently no NVM technology that can meet the low-latency and high-frequency requirements of such memories. In mainstream computer systems, NVMs are primarily installed as secondary storage in the form of SSDs (solid-state drives) or HDDs (hard disk drives), or as storage class memory to bridge the gap in latency and capacity between main memory and secondary storage. Consequently, the upper layers of the memory hierarchy have not yet achieved nonvolatility.



(a) Sequence of normally-off computing (NOF)



(b) Sequence of nonvolatile power gating (NVPG)

Figure 1.1. Time variation of power dissipation in two different energy reduction schemes with nonvolatile memory and power gating.

1.2.1 Normally-off computing

The mainstream energy saving scheme using NVM is based on Normally-off Computing (NOF) described in (Fig. 1.1 (a)). In NOF, the power supply to the memory is basically kept off while the data are always retained in nonvolatile elements, and it is only turned on when requests for memory read-/write access occur. With this approach, increasing memory access during runtime can significantly amplify energy loss. This is because the energy required for 'read/write' access to non-volatile elements (specifically referred to as 'restore/store') is significantly higher compared to electrical read/write in volatile memory.

For instance, in the case of STT-MTJ, storing refers to the action that produces the physical change of reversing the magnetization direction of the ferromagnet in the MTJ with the effect of spin-transfer torque (STT) incurred by an applied current. Restoring involves sensing the resistance of the MTJ, which varies depending on the direction of magnetization in the free layer of the MTJ. Therefore, with high-frequency memory access, the overhead of accessing nonvolatile elements might exceed the energy-saving effects of power gating. Moreover, the time required for store and restore operations is usually longer than that for read and write operations, making it challenging to meet high-frequency access demands. Finally, various types of nonvolatile elements have a limit on the number of read/write cycles, and high-frequency access demands can lead to durability issues.

1.2.2 Nonvolatile power gating

The other type of PG architecture is nonvolatile power gating (NVPG) [16]. NVPG assumes a scenario where NVM is used in intermittent operation applications with a normal operation period and an idle period, which are typical in edge computing. During the normal operation period, the application is executed, and only volatile read/write is performed during this period. Then, before and after a certain duration of idle period, store/restore is performed to retain the data in NVM (Fig. 1.1 (b)) and to recover the data to be used in normal operation. In contrast to NOF, NVPG has many advantages such as (a) reducing the number of unnecessary store/restore operations, (b) almost no increase in read/write latency during normal operation, and (c) no need to worry about endurance issues because of the number and the frequency of store/restores is significantly reduced.

1.3 Challenges of NVPG in real systems

In edge computing, devices are typically optimized for performance and energy efficiency specific to target applications. Thus, in the design phase, the system designers must analyze the characteristics of the assumed applications and choose the most suitable technologies to realize them. Although NVPG, as mentioned previously, introduces overheads such as storing and restoring to NVM along with energy savings through power gating, a breakeven analysis is essential to determine whether the overhead is less significant compared to the energy savings to decide on incorporating NVPG into a system.

Breakeven analysis requires a useful analytical model estimating energy in various application scenarios in edge computing. When considering NVPG overhead, the NVM itself must also be taken into account, meaning that it has a larger circuit footprint for the added nonvolatile functionality and, consequently, a larger leakage current during the active period compared to regular volatile memory. In NVPG, the longer the idle time, the greater the effect of PG, but if the normal operation time is long, the overhead energy from this increased leakage current due to nonvolatilization can become significant. Therefore, the breakeven analysis for NVPG must consider the balance between the lengths of normal operation periods and idle periods in intermittent operation applications.

In this research, we focus on nonvolatile flip-flops (NVFFs) utilizing STT-MTJ for NVM in NVPG. In STT-MTJ-based NVFF, efforts have been made to reduce store energy to minimize the overhead in the NVPG. In particular, Verify-and-Retryable NVFF (VR-NVFF) was proposed with a two-step store (TSS) control that can significantly reduce the store energy. TSS control is a store energy reduction method that considers the analog-like behavior of variability inherent in MTJ devices. However, the energy reduction effect of the TSS control has not been sufficiently evaluated or formulated for its full utilization in practical applications. Moreover, the VR-NVFF cell has an increased number of transistors as a result of the addition of unique store control circuitry, leading to a higher leakage current, whose negative effects also need to be considered.

1.4 Scope of this thesis and contributions

The main focus of this thesis is to address the challenges of NVPG in edge computing and to help system designers analyze the target applications and select appropriate technologies during the design phase. The thesis primarily revolves around the analysis of NVFF utilizing STT-MTJs, targeting the reduction of store energy in NVPG systems. The research conducted can

be broadly categorized into three main areas and their contributions are as follows:

1. **Real Chip Evaluation:** The first important contribution of this study is the empirical evaluation of VR-NVFF implemented in a 40 nm MTJ/CMOS hybrid process. This evaluation demonstrated the energy reduction effect of the TSS approach in VR-NVFF. The results also show how the energy reduction effect varies under different conditions and provide valuable insight into the importance of employing the TSS control under conditions where the energy reduction effect can be maximized. (Chapter 4)

2. **Energy Model Proposal:** Another important aspect of this study is the development of an energy model that integrates the variability of the MTJ switching delay times. This model, which assumes that the variation follows a normal distribution, has been shown to be useful in the search for optimal conditions that minimize the store energy of the TSS control. The model is also specifically designed for estimating energy consumption for intermittent operation applications, which are common in edge computing, and is applicable to a wide variety of intermittent operation conditions. (Chapter 5)

3. **Break-Even Analysis for NVPG:** Using the developed energy model, a break-even analysis methodology for NVPG applications is proposed. This analysis allows system designers to simulate and compare the energy consumption of different FF technologies, including volatile FF, balloon-type RFF, and MTJ-based NVFFs. As a result, it is now possible to identify the optimal technology to minimize energy consumption in various intermittent operation applications, a critical step in system design and development. (Chapter 6)

1.5 Structure of this thesis

The remainder of this thesis is structured as shown in Fig. 1.2. Chapter 2 explains the basics of CMOS VLSI power consumption and introduces characteristics and several examples of NVM for NVPG. Chapter 3 reviews some related work on the evaluation of NVFFs for NVPG to articulate motivation to build measurement based energy model for NVPG in this study. The measured results of the VR-NVFF implemented chips is evaluated in Chapter 4, which is a necessary step for the basis of the model building. Chapter 5 proposes the energy model for intermittent operation applications with NVPG using VR-NVFF as well as the models for other alternative technologies. Breakeven

analysis between several NVPG technologies using the proposed models to help system designers make an informed decision to minimize energy consumption is demonstrated in Chapter 6. Finally, Chapter 7 summarizes this thesis.

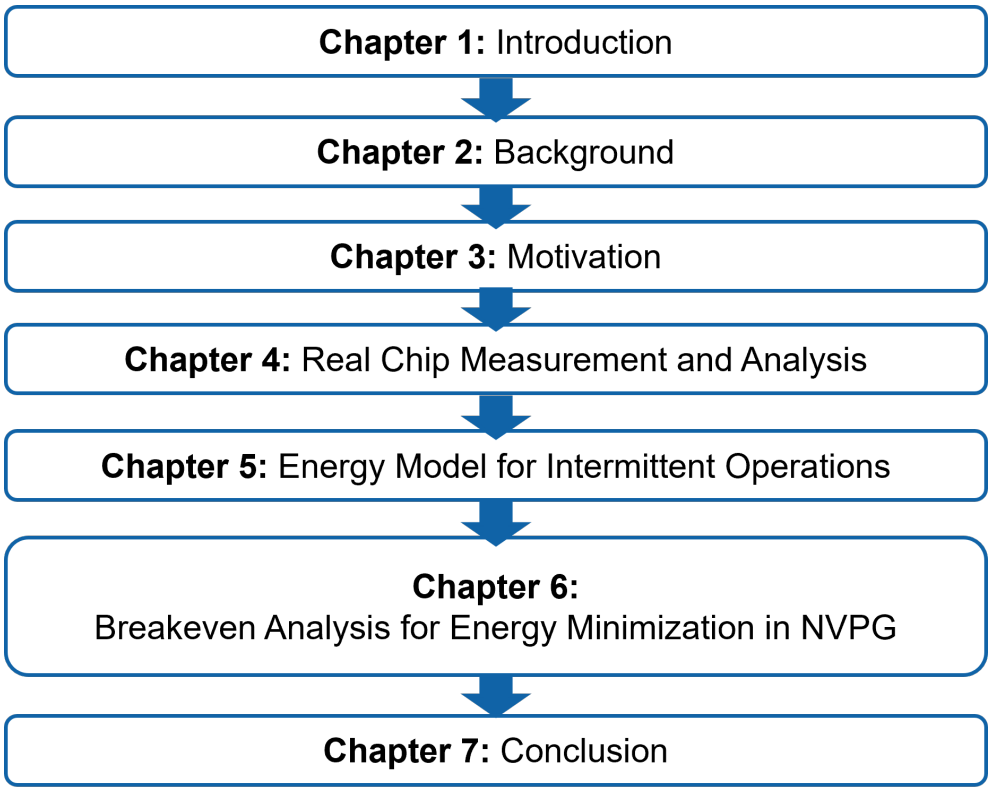


Figure 1.2. Structure of this thesis.

2

Background

In this chapter, the background knowledge related to this work, including an overview of power consumption in CMOS circuits and NVM, which is an important technology for realizing NVPG are presented.

2.1 Power dissipation in CMOS circuits

For the study of energy reduction in CMOS VLSI, it is necessary to understand the overview of power consumption in CMOS circuits. The power consumption in CMOS circuits is generally classified into dynamic power P_{dynamic} and static power P_{static}. Therefore, the total power consumption in CMOS circuits is expressed by the following equation.

$$P_{\text{total}} = P_{\text{dynamic}} + P_{\text{static}} \tag{2.1}$$

2.1.1 Dynamic power

Dynamic power is the power consumed while the signal is switching. The dynamic component of power consumption is mainly the switching power $P_{\text{switching}}$ consumed by charging and discharging the circuit when the gate input of the transistor switches from 0 to 1 or from 1 to 0, and is expressed by Equation (2.2).

$$P_{\text{switching}} = \alpha C f V_{\text{DD}}^2 \tag{2.2}$$

where α is the activity factor, C is the load capacitance, f is the clock frequency and V_{DD} is the supply voltage.

The activity factor is the probability that a circuit node transitions from 0 to 1. Since the clock rises and falls once per cycle, the activity factor for a clock signal is $\alpha = 1$. Since most data transitions occur only once per cycle, the maximum activity factor is 0.5.

Another component of dynamic power P_{dynamic} is the short-circuit current that occurs while both pMOS and nMOS are partially ON at the moment of switching. Several studies have proposed estimation formulas for short-circuit current [17, 18], but usually P_{dynamic} is dominated by $P_{\text{switching}}$. Short-circuit power strongly influenced by the the ratio $v = V_{\text{th}} \,/\, V_{\text{DD}}$. When $v > 0.5$, the short-circuit current is completely eliminated [19].

2.1.2 Static power

Static power is the power consumed by the leakage current that occurs even when the transistor is nominally ON. The static component of power consumption is further classified into four components, depending on the location where the leak occurs, 1) subthreshold leakage P_{sub}, 2) gate leakage P_{gate}, 3) junction leakage P_{jun}, and 4) gate induced drain leakage (GIDL) P_{GIDL}. Therefore, the total power consumption in CMOS circuits is expressed by the following equation, also illustrated in Fig. 2.1.

$$P_{\text{static}} = \Big(I_{\text{sub}} + I_{\text{gate}} + I_{\text{junc}} + I_{\text{GIDL}}\Big)V_{\text{DD}} \tag{2.3}$$

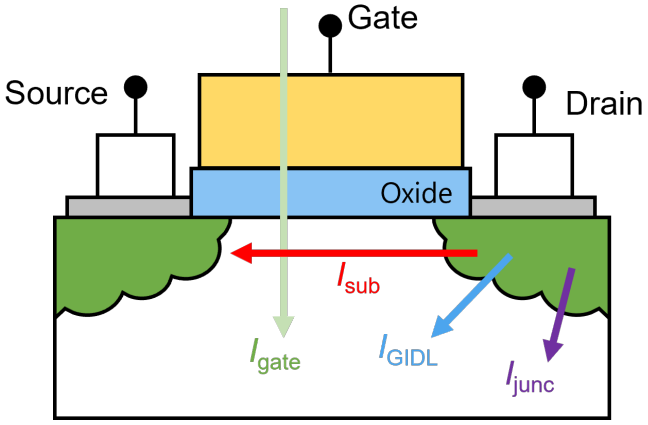


Figure 2.1. Schematic of leakage current in MOS transistor.

(a) Subthreshold leakage

Subthreshold leakage current is the leakage current flowing from the drain to the source when the gate voltage is below the threshold voltage, and its characteristics is denoted by the following equation.

$$I_{\text{sub}} = I_{\text{off}} 10^{\frac{V_{\text{gs}}+\eta(V_{\text{ds}}-V_{\text{DD}})-k_\gamma V_{\text{sb}}}{S}} \tag{2.4}$$

where I_{off} is the subthreshold current at $V_{\text{gs}} = 0$ and $V_{\text{ds}} = V_{\text{DD}}$ while V_{gs}, V_{ds}, and V_{sb} are the gate-source, drain-source, and source-body voltages, respectively. η is a coefficient of the drain-induced barrier lowering, k_γ is the body effect parameter, and S is the subthreshold slope [20]. Subthreshold leakage increases exponentially with the decrease in the threshold voltage V_{th} of CMOS transistors. Since V_{th} needs to be lowered to operate the IC at high speed, the operating speed of the IC and the subthreshold leakage current are in a trade-off relationship.

(b) Gate leakage

Gate leakage current is the leakage current flowing between the gate and the substrate, gate and source, and gate and drain. Since there is a gate insulating film, the gate leakage current should not flow. However, as the thickness of the gate insulating film is less than 2 nm (the thickness of 5-6 atoms) due to the miniaturization of semiconductor devices, the gate leakage current flows from the gate to the substrate due to the quantum tunneling effect.

$$I_{\text{gate}} = WA \left(\frac{V_{\text{DD}}}{t_{\text{ox}}} \right) e^{-B \frac{t_{\text{ox}}}{V_{\text{DD}}}} \tag{2.5}$$

where W is the gate width, A and B are parameters depending on the CMOS process technology, and t_{ox} is the thickness of the gate oxide. As the equation suggests, I_{gate} increases exponentially as the gate dielectric t_{ox} becomes thinner. This effect cannot be ignored as the LSI manufacturing processes become finer and finer, as will be discussed later in the section.

(c) Junction leakage

Junction leakage I_{junc}, flowing through P-N junctions between the drain and the substrate and the source and the substrate, occurs when electrons pass through the depletion layer because the n-type semiconductor area becomes smaller due to miniaturization, but at the same time the depletion layers between the source and substrate and between the drain and substrate become smaller.

(d) GIDL

GIDL I_{GIDL} is the current flowing from the drain to the substrate when a high electric field is applied to the drain end under the gate electrode. This effect is most significant when the drain is at high voltage and the gate is at a low voltage. The GIDL current is proportional to the gate-drain overlap area and therefore to the transistor width. The GIDL current increases rapidly with the drain-gate voltage because it is a strong function of the electric field. In particular, when both the drain and the source are heavily doped, band-to-band tunneling (BTBT) accounts for most junction leakage [21].

2.1.3 Increased leakage power due to miniaturization

The performance of LSI has been improved by miniaturization of CMOS process technology, which has led to faster operation, lower power consumption, and higher integration. However, as miniaturization progresses, the power consumption due to the leakage current has become significant [21,22], although it does not contribute to the execution of applications. This is primarily due to the increase in subthreshold leakage resulting from a lower threshold voltage and the shorter channel length and the increased gate leakage resulting from thinner gate oxide. The trend projections of transistor physical dimensions and device power consumption based on the ITRS (International Technology Roadmap for Semiconductors) as of 2002 are shown in Fig. 2.2. All values are normalized to the 2001 values. The exponential increase in leakage current was expected to cause the static power consumption to exceed the dynamic component of the power consumption unless effective measures were taken to reduce the leakage power [5, 21]. Recent state-of-the-art studies [23–25] have actually shown that the leakage current as a percentage of the total chip is still significant, at 10-30 % or more.

The consumption of unnecessary static power that does not contribute to the execution of applications is a major problem, especially for devices with limited power capacity, such as mobile devices and IoT devices. In particular, in intermittent operation applications that are common in edge computing, it is quite possible that the standby power consumption is greater than the energy required to execute the application. Therefore, reducing leakage current during the idle period of intermittent operation applications is an effective approach to improve the energy efficiency of the VLSI system.

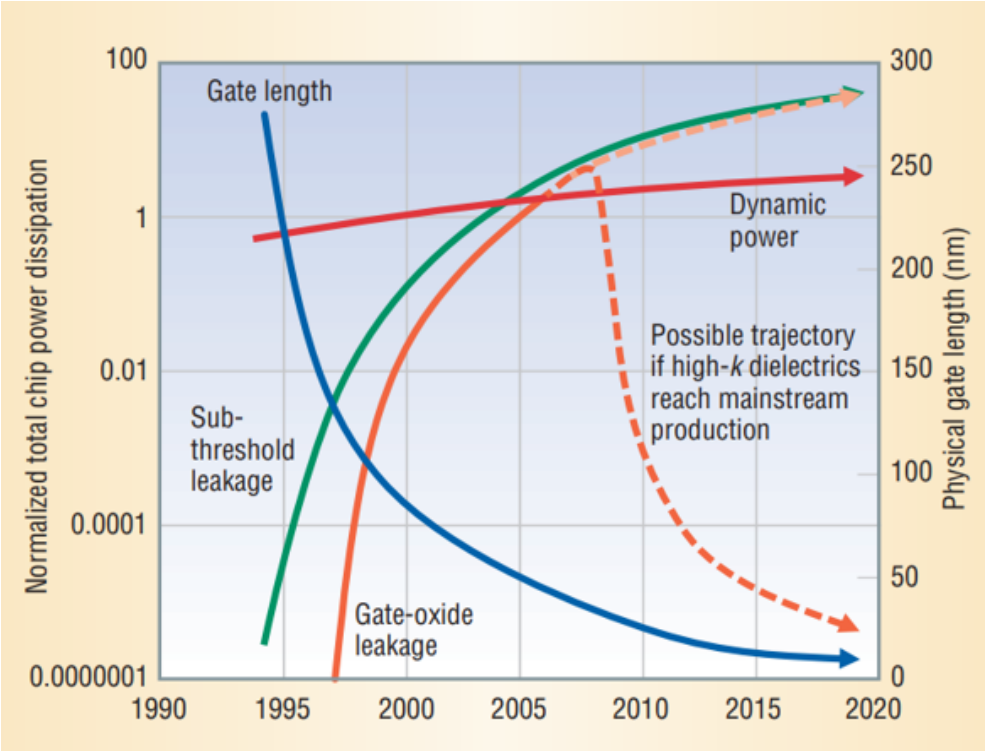


Figure 2.2. Dynamic and static power trend projections based on the ITRS as of 2002 [5].

2.2 Emerging nonvolatile memory

NVM is a type of memories that is able to retain the stored information even when the power is turned off. Established NVMs include HDD (Hard Disk Drive), flash memory, and EEPROM (Electrically Erasable Programmable Read-Only Memory). These memories are widely used for long-term data storage, and HDDs in particular are suitable for storing large amounts of data at low cost. Flash memory provides faster access speed and physical durability, and is widely used in portable storage devices such as USB drives and SSDs. EEPROM is suitable for small-scale data storage, allows data to be rewritten, and is used in programmable logic devices and small electronic devices.

On the other hand, emerging CMOS-compatible NVM or embedded NVM, i.e., PCRAM, RRAM, and MRAM, which merges NVM into CMOS circuitry are recently come to market or nearing to commercialization. According to a recent forecast report on emerging NVM by Yole [26], a MEMS-related market research firm, the market size related to these new technologies is expected to reach the equivalent of $2.7 billion by 2028. The remainder of this section provides an overview of emerging NVM technologies that have attracted attention in recent years, leaving the overview of these established NVMs to other excellent literature.

2.2.1 PCRAM

PCRAM (phase-change memory) is a memory technology that stores data by switching a GST (Germanium-Antimony-Tellurium) film between two distinct states: amorphous (reset, which is high resistance) and crystalline (set, which is low resistance). This process leverages the significant difference in resistivity between the amorphous and crystalline phases of phase-change materials. Electrical currents are applied to repeatedly toggle the material between these two phases.

PCRAM is gaining attention as a potential alternative to DRAM. Its primary advantages other than nonvolatility are resilience to soft errors, which are transient errors not caused by permanent damage, low read latency and good scalability, making it a promising candidate for various memory applications [27]. However, there are certain drawbacks when compared to DRAM: PCRAM generally consumes more power, suffers from longer write latencies, and has a shorter overall lifespan.

A well-known example of PCRAM products is Intel's Optane series [28], which adopts 3D XPoint technology developed by Intel and Micron. Intel released Optane DC Persistent Memory and Optane DC SSD as Storage Class Memories with intermediate characteristics between DRAMs and NAND flash

memories in terms of the capacity and the latency. However, Intel was unable to achieve commercial success due to complex reasons such as ecosystem building and profitability issues, and withdrew from the Optane business in 2022, resulting in the disappearance of companies that commercially produce PCRAM today.

2.2.2 ReRAM

ReRAM (resistive RAM) operates by changing the resistance of a specially formulated solid dielectric material, known as a memristor – a contraction of "memory resistor." Its structure is based on a simple three-layer formation: a top electrode, a switching medium, and a bottom electrode. The resistance switching mechanism within ReRAM is activated when a voltage is applied between these electrodes, leading to the formation of a filament in the switching material. This unique structure allows for data storage and retrieval by varying resistances under different voltage applications.

ReRAM is increasingly recognized as a promising memory technology, particularly suitable for System on Chip applications due to its fast read/write performance, low power consumption and make it one of the most promising memories available. Additionally, ReRAM offers several advantages over other memory technologies like DRAM, PCM, and MRAM, including better water resistance due to its internal structural simplicity and material stability. However, it faces challenges such as high cost and difficulties in the etching process, which hinder its widespread adoption in various IoT applications. ReRAM is also byte-addressable, distinguishing it from flash memory, and boasts higher density and greater endurance.

2.2.3 MRAM

MRAM utilizes the magneto-resistive effect, which uses the magnetic states of magnetic materials, which change in response to magnetic fields, to store data. In MRAM's evolution, different types of MTJs have been developed, each with unique mechanisms and advantages.

Toggle-MTJ, the early form of MRAM, uses an external magnetic field to change the magnetic orientation in the MTJ. However, its reliance on external magnetic fields often requires complex circuitry, which can limit memory cell density and scalability. STT-MTJ is a more advanced approach, employing the spin of electrons to alter magnetic directions, enabling efficient and denser memory architectures. STT-MTJ is known for its energy efficiency and scalability, making it suitable for a wide range of applications. SOT-MTJ (Spin-Orbit Torque-MTJ) is a newer variant that uses spin-orbit torque for

switching, offering prospects of faster speeds and potentially lower energy consumption. Its writing mechanism employs in-plane induced currents to switch the state of the MTJ without passing through the junction, enhancing reliability. However, SOT-MTJs are still in the early stages of research and face challenges such as the higher current density required for switching, which requires larger transistors than STT devices, affecting efficiency and scalability.

In near future, MRAM is expected to dominate the embedded NVM market over RRAM according to Yole [26]. In particular, STT-MTJ-based MRAM is expected to further expanding its range of applications and offering the greatest revenue potential compared to RRAM and PCM (MRAM revenue equals RRAM and PCRAM combined in total revenue in 2028).

(a) STT-MTJ

STT-MTJ-based MRAM has recently entered the commercial production stage. Various foundries announce the readiness of embedded MRAM (TSMC, GlobalFoundries, Samsung) for production. For embedded applications, STT-MRAM possesses non-volatility, high-endurance, scalability, low power, and fewer masks than the embedded Flash. It also provides great area savings and lower leakage compared with SRAM [29].

A STT-MRAM cell consists of MTJ with two ferromagnetic layers, i.e., a free layer and a pinned or reference layer sandwiching a tunnel barrier. As shown in Fig. 2.3, the MTJ element changes between two states: (a) the magnetization directions of the two ferromagnetic layers are parallel and (b) the magnetization directions of the two ferromagnetic layers are anti-parallel, and the resistance of the MTJ element is in a low resistance state and a high resistance state, respectively.

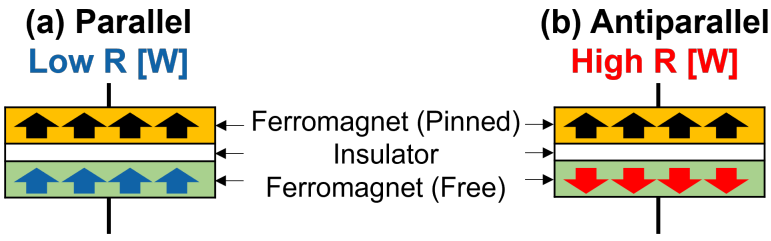


Figure 2.3. Schematic diagram of STT-MTJ consisting of insulator oxide sandwiched between two ferromagnetic layers (pinned layer and free layer): (a) low resistance state of MTJ element with parallel magnetization of two ferromagnetic layers, (b) high resistance state of MTJ element with anti-parallel magnetization of two ferromagnetic layers.

In essence, the magnetization switching of the ferromagnetic layer occurs due to the applied magnetic field. In STT-MTJ, the current flowing through the ferromagnetic layer causes the magnetization to switch by applying a torque, called spin-transfer torque (STT) that reverses the axis of rotation of the precessing electron spin. This method is smaller and more salable than the Toggle-MTJ method, which requires devices such as a coil or dedicated wiring to generate a magnetic field to change the direction of magnetization externally. It also has the advantage of low power consumption because the energy required for magnetization reversal is small.

Moreover, STT-MTJ comes in two types: in-plane and perpendicular, distinguished by the orientation of magnetization within the junction. Compared with its in-plane counterpart, perpendicular STT-MRAM has higher density, lower switching current, and is easier to control in large scale manufacturing. In the in-plane STT-MTJ, the magnetization lies within the plane of the junction. This structure, common in earlier STT-MTJ designs, allows data recording by altering the direction of magnetization within the plane via an electric current. On the other hand, perpendicular STT-MTJ features magnetization that is oriented perpendicular to the plane of the junction as shown in Fig. 2.3. This perpendicular alignment offers higher thermal stability and more efficient writing. Perpendicular STT-RAM has better write energy, endurance, and performance than traditional NVM such as eFlash. It has a comparable read speed than that of SRAM, and its single-cell structure presents much higher density than SRAM.

2.3 Nonvolatile memory for NVPG

The requirements for NVM to realize NVPG are summarized, and several examples of NVM that satisfy these requirements are given.

2.3.1 Requirements for NVM utilized in NVPG

There are two requirements for NVM to realize NVPG, one in terms of configuration and the other in terms of functionality. [16] as depicted in Fig. 2.4.

The first requirement is the configuration, which basically requires bistable circuits, nonvolatile elements, and the circuits to control the nonvolatile element. Examples of bistable circuits include latches and flip-flops, and examples of nonvolatile elements include MTJs for MRAM and phase change materials for PCRAM.

The second requirement is that it can switch between two operating modes: “normal operation mode” and “nonvolatile control mode”. The application is executed in normal operation mode, at which time the memory works as a

normal latch or flip-flop and no access is performed to the nonvolatile elements. Therefore, no energy consumption or delay occurs caused by NVM access. Before and after power gating, the operating mode switches to nonvolatile control mode, and the control circuit performs store and restore operation to realize the nonvolatile retains. It should be noted that the static power consumption during normal operation rises because of the additional control circuit for nonvolatile elements.

In the rest of this section, several examples of NVM that satisfy these two requirements are introduced.

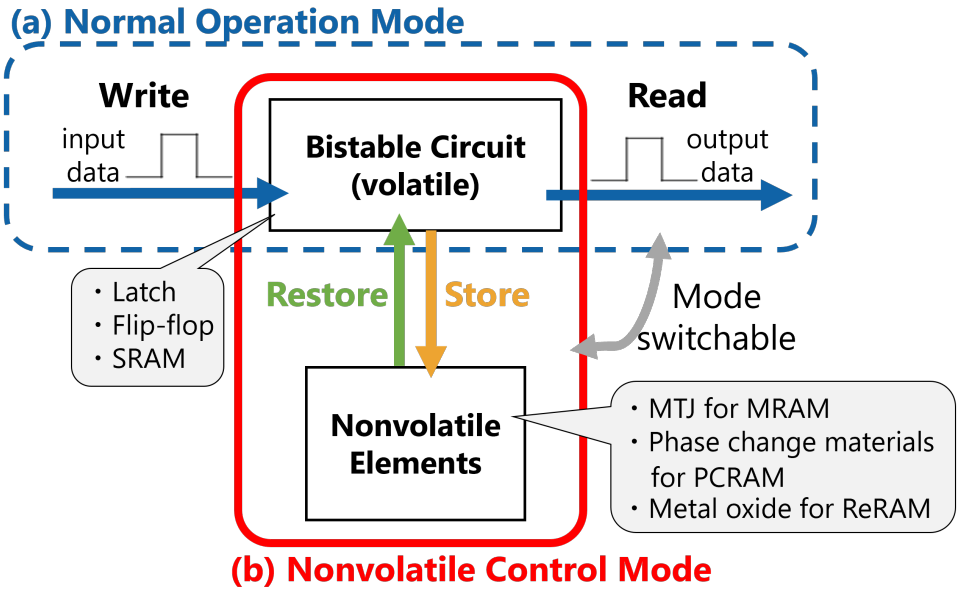


Figure 2.4. Schematic diagram of nonvolatile memory requirements for NVPG.

2.3.2 Pseudo-spin-MOSFET NVFF

Pseudo-spin-MOSFET nonvolatile flip-flop (PSM-NVFF), proposed by Yamamoto et al. [30, 31], is a NVFF for NVPG utilizing STT-MTJ. A cell structure of PSM-NVFF is based on an ordinary volatile leader-follower type FF shown in Fig. 2.5. Additional three transistors and two MTJs constitute PSM-NVFF shown in Fig. 2.6.

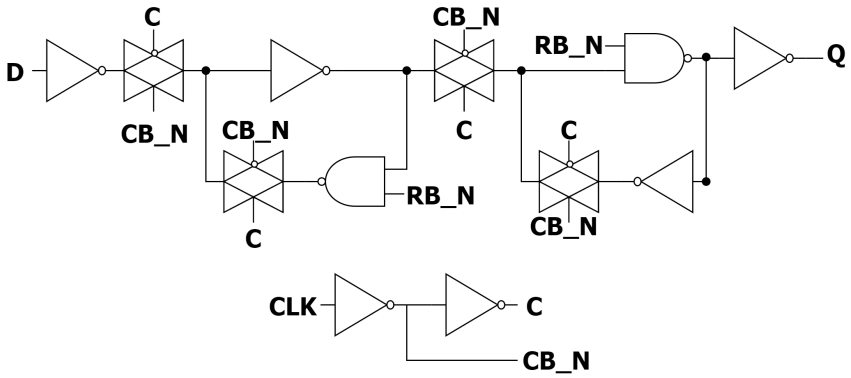


Figure 2.5. Cell structure of ordinary volatile leader-follower FF: VFF.

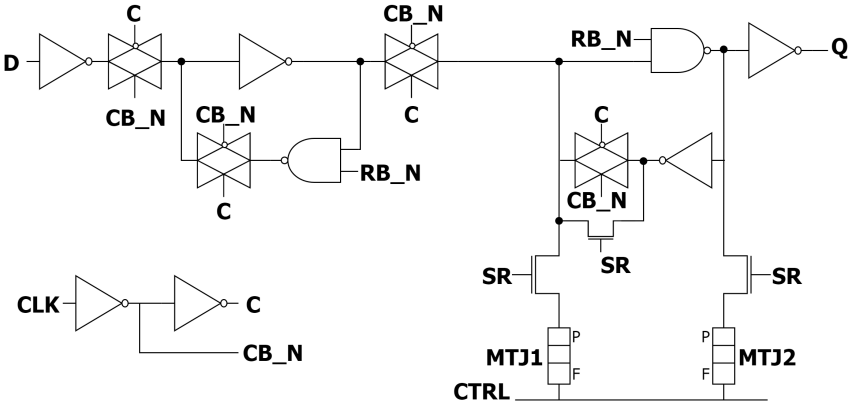
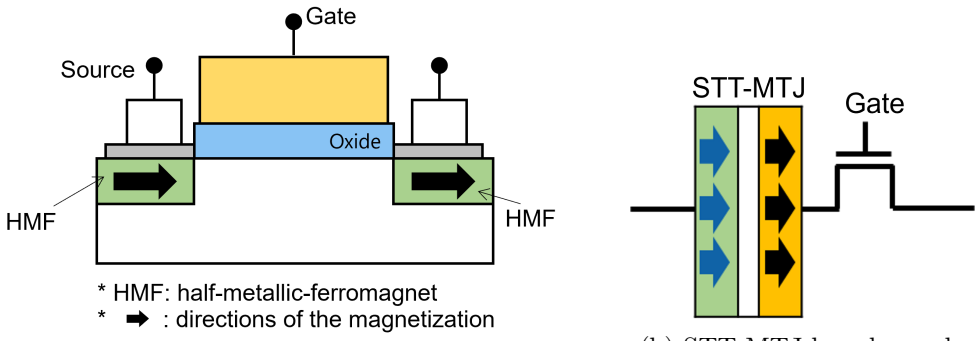


Figure 2.6. Cell structure of pseudo-spin-MOSFET NVFF: PSM-NVFF.

The combination of one MTJ element and one CMOS transistor is a circuit technology called pseudo-spin-MOSFET [32], illustrates in Fig. 2.7 (b), which perform the same operation as spin-MOSFET [33], shown in Fig. 2.7 (a). Spin-MOSFET has a structure in which a ferromagnetic electrode is placed on the source and drain of the conventional MOSFET, and the direction of magnetization of the ferromagnetic electrode controls the ease of current flow between the source and drain, behaves as a transistor, and functions as a magnetic resistance element as well. Although spin-MOSFET has been studied actively, the performance of current devices is far from practical applications [29, 34, 35]. PSM has benefited from the research results on MRAM and has realized the ideal operation of spin-MOSFET, which is a realistic and available technology today.



(a) Device structure of spin-MOSFET.

(b) STT-MTJ-based pseudo-spin-MOSFET.

Figure 2.7. Spin-MOSFET and pseudo-spin-MOSFET.

PSM-NVFF performs as a normal FF during normal operation mode, and during nonvolatile control mode, the selection transistor are controlled by the SR signal are controlled to store data in the MTJ or restore data from the MTJ. During normal operation mode, the follower latch and the MTJ are electrically separated by the selection transistors, and therefore the influence on the operation as a normal FF is insignificant. Depending on the data of the follower latch, store operation changes the state of two MTJ to the low resistance state and the high resistance state, respectively. In the restore operation, the state of the follower latch is read out by the difference in the current caused by the difference in the resistance between the two MTJs. NVPG is realized by performing the store operation before PG and the restore operation after PG.

However, PSM-NVFF has the following disadvantages due to its structure. Since the follower latch and the MTJ are connected by a simple PSM, there is a risk of "latch destruction" in which the current flows back and rewrites the data of the follower latch during the store operation, and to avoid this, it is necessary to increase the size of the store control transistor. Additionally, since the store and restore operations are controlled by the same transistors, an unnecessarily large current flows during the restore operation, resulting in a waste of energy.

2.3.3 Split Store/Restore NVFF

Split Store/Restore NVFF, proposed by Kudo et al. [36], is a circuit that improves the robustness during the store operation and optimizes the restore current and reduces the cell area compared to PSM-NVFF. As for the circuit structure as shown in Fig. 2.8, while PSM-NVFF shared the paths for store

and restore operations, SSR-NVFF separates the paths for each operation, allowing the size of the transistors to be optimized according to the amount of current required for store and restore operations. Inserted inverters into the store path prevent the current flow that causes the latch destruction.

Although the number of transistors required for SSR-NVFF is larger than that of PSM-NVFF, it is reported that the area of the entire circuit can be reduced by 45% by effectively reducing the size of each transistor [36]. In this way, SSR-NVFF not only reduces the energy consumption during store and restore operations but also achieves a reduction in the active leakage current by optimizing the circuit.

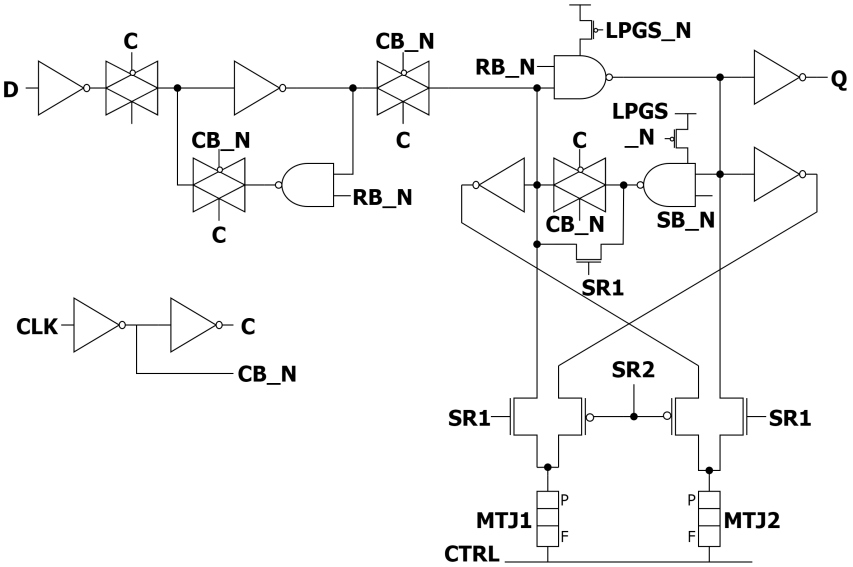


Figure 2.8. Cell structure of Split Store/Restore NVFF: SSR-NVFF.

2.3.4 Verify-and-Retryable NVFF

VR-NVFF, as depicted in Fig. 2.9 and proposed in the work by Usami et al. [37,38], is a further improved NVFF that introduces two notable features: a) the data aware store (DAS) function, which stores data in the MTJ only if the data in the MTJ differs from the data to be stored in the FF, and b) the energy-saving store method, the two-step store (TSS) control.

VR-NVFF consists of a common master latch and a follower latch, as well as two MTJs. It is distinguished from SSR-NVFF by its balloon latch and independent paths from the MTJs for verification and retry features including a XOR gate. A cell of VR-NVFF is controlled by a total of 10 signal lines (8 NVFF control lines, a PG control line, and a clock gating control line).

In the following, the DAS function and the TSS control will be discussed. These functionalities are represented by the words 'Verify' and 'Retry' in the name VR-NVFF, respectively.

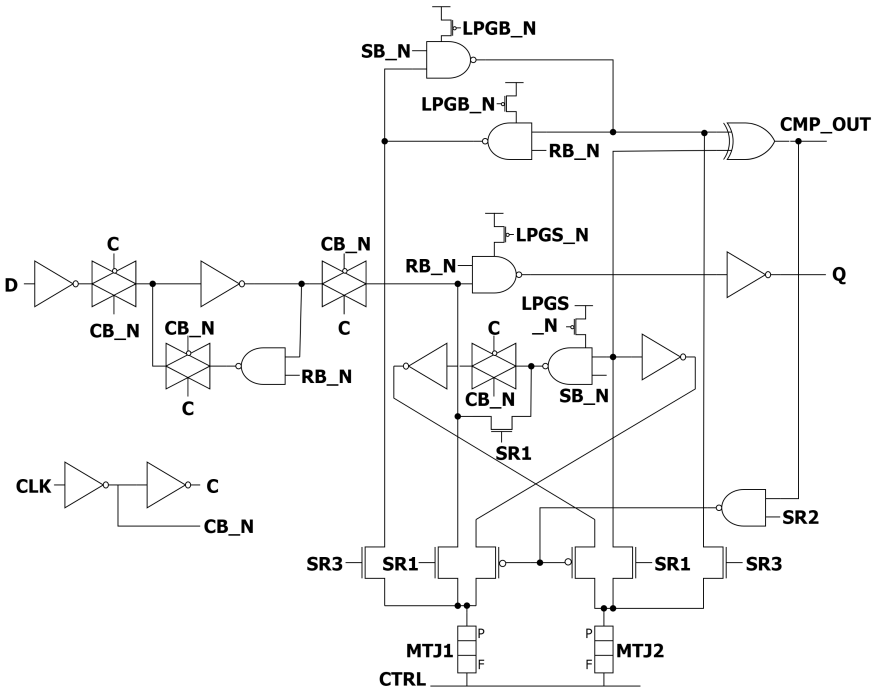


Figure 2.9. Cell structure of Verify-and-Retryable NVFF: VR-NVFF.

(a) Data Aware Store

The first key feature of VR-NVFF cell is the verification operation that realizes the DAS function. In PSM-NVFF and SSR-NVFF, all data in the follower latch are written to the MTJs during the store operation. However, if the data in the follower latch and the data retained in the pair of MTJs are equal, there is no need to store the data in the MTJs. The DAS function avoids such unnecessary store operations by checking the data in the follower latch and the data retained in the MTJs, thereby reducing the store energy.

DAS function is realized by the verification operation that reads out the data from the MTJ, moves it into the balloon latch, and compares the data in the balloon latch with the data in the follower latch using the XOR gate. When the two data from two latches are the same, the output (CMP_OUT) of the XOR gate becomes "0", and then the transistors TR1 and TR2 are forced to turn off by CMP_OUT, and consequently the current to the MTJs

is cut off. This series of operations is called verify operation. In this manner, the VR-NVFF can validate whether or not the store operation is required in bit-level units, consequently avoiding unnecessary power consumption.

The effect of the DAS function depends on the application, but Kudo et al. [39] reported that the evaluation using the ISCAS'89 benchmark as an application achieved 72 - 86 % energy reduction compared to the PSM-NVFF.

(b) Two-step Store

The other unique energy-saving policy of VR-NVFF is the TSS control.

The main idea behind the TSS control is to consider the fluctuations in the switching delay time, namely the minimum time required to change the magnetization. These fluctuations are caused by the process variation of the MTJ/CMOS hybrid process, variations in local supply voltage, and the inherent stochastic nature of MTJs.

Due to the process variation in CMOS and MTJ devices and the physical characteristics of MTJ devices, the time (or energy) required for the magnetization switching of the STT-MTJ is not constant, but is known to vary according to some probability distribution [40–42] as depicted by the blue line in Fig. 2.10, showing that most of the MTJs require a short time to switch the magnetization, but some of them require a long time. In memory systems, multiple NVFFs are usually incorporated as a single unit, like the SD in NVCMA/MC, and the store control is performed collectively. Therefore, a sufficiently long store operation is required assuming the MTJ element of the worst case that takes the longest time to switch the magnetization. However, this method also causes a significant waste of energy because a long store operation is performed even for MTJs that require a short time to switch the magnetization. The orange line in Fig. 2.10 shows the pass rate (PR), given by Equation (2.6), which indicates the percentage of NVFFs in the SD that have successfully completed the store operation.

$$PR = \frac{\#\,of\ successfully\ stored\ NVFFs}{N_{\mathrm{store}}} \tag{2.6}$$

where N_{store} is the number of NVFFs to tried to be stored. It is worth noting that the switching delay time variation (blue line) and PR (orange line) are related by the probability density function and cumulative distribution function of a certain probability distribution.

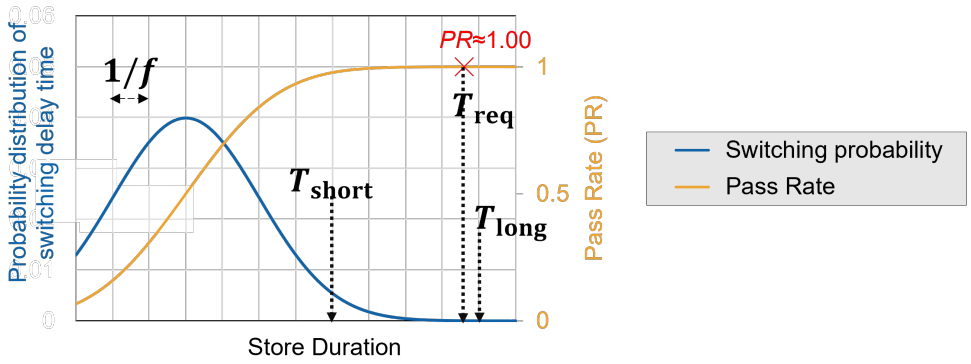


Figure 2.10. Variation in switching delay time of MTJ-based NVFFs and pass rate.

In the TSS control, in order to reduce such waste, the store is performed in two stages according to the flow shown in Fig. 2.11.

Follwed by the first short store operation, the verify operation is performed to check the success or failure of the store operation on each NVFF. The second trial (retry) is performed only for the NVFFs that failed in the first try with sufficiently long time. This methodology significantly reduces the energy waste by avoiding the extra current flow to the major MTJs that successfully switched the magnetization in a short time.

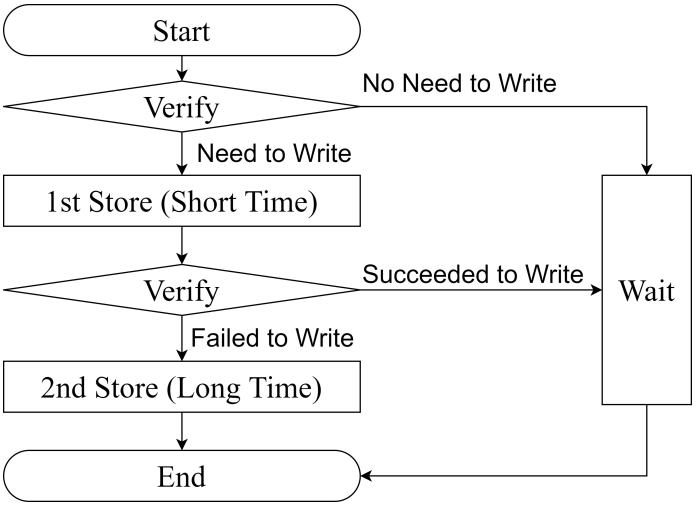


Figure 2.11. TSS control flow of VR-NVFF.

Fig. 2.12 shows a timing chart of an NVFF cell that successfully switches the MTJs in the first short store during the TSS control. Since the NVFF

cell confirms with the verify operation that the MTJs successfully complete the store operation, the unnecessary current does not flow to the MTJs of this cell during the second store operation. Note that the "Store" operation takes three steps; two MTJs (MTJ1 and MTJ2) are stored one at a time, and subsequently a one-clock slack time is inserted for a safe operation. T_{short} and T_{long} are the periods of time for applying current to one of the pair of MTJs at short and long store operations, respectively.

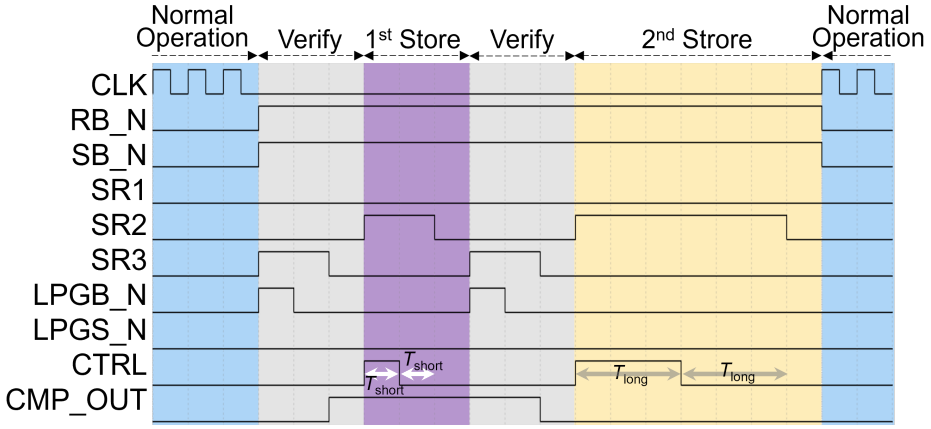


Figure 2.12. Timing diagram of the control signals of VR-NVFF under the TSS control.

T_{short} and T_{long} are multiples of the inverse of the control signal frequency $(1/f)$ in real systems, and the following relationship expressed in Equation (2.7) holds as also illustrates in Fig. 2.10.

$$T_{\mathrm{long}} \geq T_{\mathrm{req}} > T_{\mathrm{short}} \tag{2.7}$$

where T_{req} represents the enough long store time that is required considering the worst-case MTJ-based NVFF cell, and is the time at which the PR saturates (ideally PR becomes 1.00 without hardware failures).

The energy-saving effect expected by the DAS function and the TSS control is illustrated in Fig. 2.13. Here, in contrast to TSS control, a method that is completed with a single set of verify and store operation is called the one-step store (OSS). In the OSS control, the store power is reduced compared to the SSR-NVFF because the current is applied only to the NVFFs that require the store operation due to the effect of the DAS function. However, the store energy is still large because a current is applied to all NVFFs to be stored for a sufficiently long time in order to complete the switching in a one-time store

operation. In the TSS control, it is expected that the total store energy can be further reduced by performing two times of DAS with different store times.

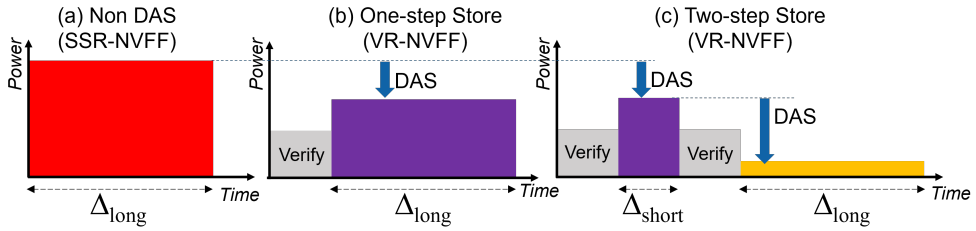


Figure 2.13. Comparison of power transition in store operation: Non DAS with SSR-NVFF vs. OSS with VR-NVFF vs. TSS with VR-NVFF.

(c) Optimal T_{short} in the TSS control

What needs to be noted before using the TSS control is that the T_{short} can greatly affect the total store energy of the TSS control. As shown in Fig. 2.14, if T_{short} is too short, it will not be able to store enough NVFFs and will result in an increase in long store energy. On the other hand, if T_{short} is too long, it will consume excessive energy on the majority of NVFFs that are flipped quickly. Therefore, T_{short} needs to be optimized to minimize the total store energy. However, to optimize T_{short}, the variability characteristics of MTJs need to be understood and even modeled so that energy savings by the TSS control can be maximized.

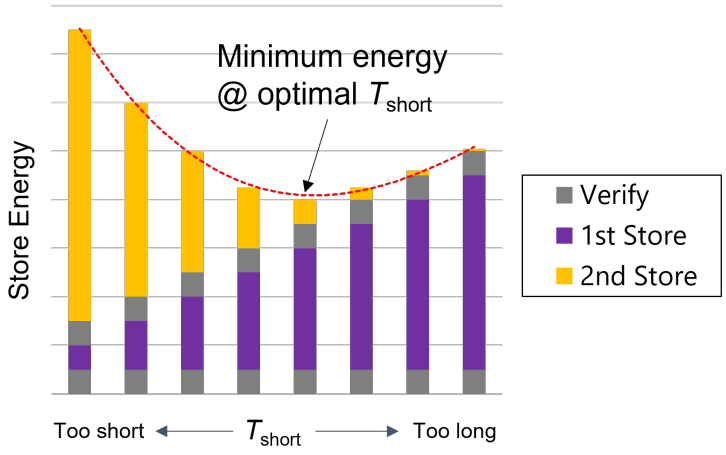


Figure 2.14. Store energy of the TSS control at different T_{short}.

2.3.5 Retention Flip-Flop [1–4]

At the end of the section, retention flip-flop (RFF), which are not technically "nonvolatile" but can retain data with ultra-low power consumption, is also mentioned as a type of memory for NVPG. As a typical RFF, the balloon-type RFF, which was first proposed by Shigematsu et al. [1], has a balloon consisting of high-V_{th} transistors added to an ordinary volatile FF, as shown in Fig. 2.15. During the normal operation period, it behaves as an ordinary leader-follower FF, but during the non-operation period, it stores data in the balloon latch and cuts off the power supply other than the balloon latch to reduce the standby leakage power. To minimize standby leakage power, the balloon latch is composed of high-V_{th} transistors, making the cell a multithreshold voltage circuit. According to the simulation results [3] with a typical corner of a 40 nm CMOS process, the balloon-type RFF can reduce the leakage power during sleep ($\tilde{1}$00 pW) to about 1/5 of the leakage power during active ($\tilde{5}$00 pW). The advantages of RFF over NVFF are that the energy required for store/restore operation is very small, so that the energy saving effect can be obtained even with short-term PG, and that the manufacturing process can be completed only with the CMOS process.

However, it should be noted that one RFF cell contains both the PG application area (FF part) and the non-application area (balloon part). To realize this, it is necessary to introduce dual V_{DD} (or multi V_{DD}) technology that uses multiple power rails [43, 44], which inevitably complicates the cell design. Furthermore, since the standby leakage power of the RFF cell is not zero, the total energy consumption increases if the idle time is longer than a certain period. Thus, RFF is suitable for short-term PG, and is in contrast to 'true' NVFFs that can reduce energy consumption as the PG time increases.

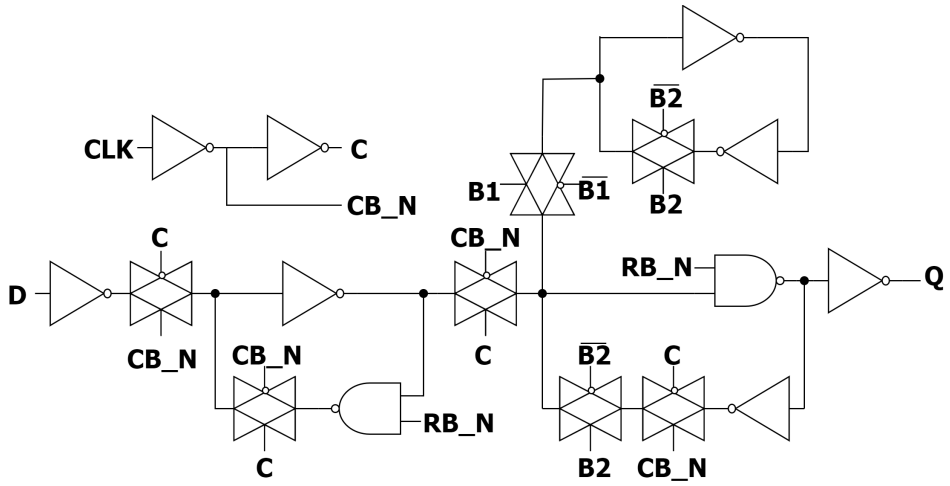


Figure 2.15. Retention FF with a balloon latch: RFF.

3

Motivation

In this chapter, we first organize the key concepts that must be considered in the design phase of energy-efficient edge devices utilizing NVPG. Following that, we review previous studies on NVFF, focusing on how NVFF has been evaluated in NVPG scenarios. We argue that existing NVFF evaluations are insufficient to make well-informed decisions about the adoption of NVFF in NVPG to minimize energy consumption in intermittent operation applications. This inadequacy has been a motivating factor in conducting this research.

3.1 Considering NVPG in edge device design

Devices in the edge environment, such as sensor nodes, wearable devices, and devices in IoT systems, are generally application-specific integrated circuits in the broad sense. They are tailored for assumed applications, predefined usage conditions, and optimization purposes. Therefore, the system design process involves selecting the most suitable technologies through comparative evaluations under conditions assumed by the target applications.

When considering the introduction of NVPG with NVFFs into your system, it is crucial to assess whether the energy reduction benefits of NVPG deployment outweigh the associated overhead. Furthermore, this assessment should be based on reliable data, especially when introducing new process technologies embedding NV elements, which vary greatly depending on the foundry and their processes.

Standard evaluation indices, for example, power consumption for each

static state, data retention capability, read/write speed, area efficiency, endurance, and so on, are often used to evaluate individual NVFFs when they are proposed. While each index is important and can be used to compare the characteristics of alternative technologies, it should be noted that each technology tends to have its own strengths and weaknesses. Therefore, the decision about which technology is the best often depends on the specific application and operating conditions. In cases where the primary objective is to minimize energy consumption, the following essential evaluations should be conducted to realize energy reduction through the NVPG with NVFFs:

- Evaluation based on measurement of fabricated NV elements. The significance of actually implementing the NVFFs on a chip and demonstrating their performance for energy reduction is paramount. Specifically, for NVFFs employing a CMOS/NV-element hybrid process, the behavior of the NV component, which plays a crucial role and possibly consumes significant energy, cannot be simulated with basic SPICE and requires complex analog models. However, the characteristics of NV elements can vary greatly depending on the manufacturer and process. Therefore, it is essential to analyze the characteristics of the fabricated chip for each process and collect data to build a model that can be used for decision making with sufficient reliability. Relying on empirical data for estimating energy allows for more precise decision making in system design and provides valuable feedback that can help refine the fabrication process.

- Breakeven analysis considering additional leakage current.

  Breakeven analysis is a calculation that weighs the total benefits against the total costs when introducing new approaches, such as NVPG in this context. Breakeven time (BET) is one of the most common indices to assess the performance of PG systems. It denotes a certain duration of standby time during which the energy savings achieved through NVPG are negated by the additional energy consumption of NVFFs. BET can serve as a criterion for deciding whether to employ the NVPG scheme. For the scheme to be energy efficient, the standby time of the target intermittent application must exceed the BET, ensuring that the reduced energy outweighs the associated overhead.

  Ideally, all overheads should be included in the breakeven analysis to facilitate an informed decision. At the NVFF cell level, the primary overheads of NVPG include the energy required for store and restore operation for NV elements and the additional leakage current during the normal operation of the intermittent application. This extra leakage, caused by circuits controlling NV elements in the cell, results in an in-

crease in energy overhead as normal operation continues. Therefore, it is essential to consider BET as a function of the normal operating time (T_{OP}).

3.2 Related work on the evaluation of NVFFs for NVPG

Various NVFFs, including those described in Chapter 2, have been proposed and evaluated [30,31,36,39,45–59].

Although most of the proposed NVFFs meet the requirements for memory in NVPG as described in Section 2.3, thus being reasonably suitable for NVPG, many of them [47,48,50,53,56–58] lack a specific evaluation in the context of NVPG. Instead of conducting breakeven analyses, these studies often focus solely on optimizing store operation, including minimizing store energy or decreasing error rate. On the other hand, studies such as [49,51,59] concentrate on developing efficient data preservation schemes in scenarios that can involve sudden power shutdowns, such as those found in energy-harvesting settings, rather than focusing primarily on energy minimization by NVPG.

Kudo et al. proposed SSR-NVFF [36] and conducted SPICE simulations to compare the characteristics with PSM-NVFF as well as another STT-MTJ-based NVFF, selectively store NVFF (SS-NVFF) [39]. The architecture of SS-NVFF cell is based on that of PSM-NVFF, with an additional latch that provides the DAS function to reduce store energy depending on the data. In other words, SS-NVFF accepts an area overhead, leading to increased leakage energy consumption, in exchange for reduced store energy. The evaluation includes various performance indices such as area, C-Q delay, dynamic energy, and store/restore energy, along with BET when compared to a non-PG scenario. The cell area of SSR-NVFF was reported to be effectively reduced by separating store/restore path, resulting in a cell size of 0.52x and 0.40x compared to PSM-NVFF and SS-NVFF, respectively. On the other hand, the BET of SS-NVFF was evaluated to be shorter (better) than that of SSR-NVFF, because of the preferable effect of the store energy reduction by the DAS function in SS-NVFF. However, this BET evaluation is not fair for SSR-NVFF, as it does not consider the increased leakage current due to the larger cell size of SS-NVFF. Moreover, when the period of normal operation is longer, the higher leakage power of SS-NVFF becomes significant. Consequently, the evaluation conducted in the study [36] failed to provide guidance in determining which NVFF or even a non-PG method is the most effective in minimizing overall energy consumption. [30,39,46,52,55] also fail to account for increased leakage power during normal operation when analyzing the BET.

[45] conducted a quantitative analysis of the effect of NVPG with PSM-NVFF. The increased leakage overhead caused by additional circuits was taken into account when analyzing the BET. Thus, BET is defined as the sum of BET_{SR}, which is the BET due to the store and restore operation, and BET_L, which is due to the static leakage current during normal operation. The overhead can be expressed as a function of the normal operation period T_{OP}. This makes it possible to estimate which intermittent operation applications can benefit from NVPG using PSM-NVFF, considering various values of T_{OP} and T_{NOP}. However, the calculation model is an estimated value using a SPICE simulator that incorporates the MTJ macro-model, and no evaluation has been performed on actually fabricated chips.

[54] fabricated a chip to evaluate an NVFF using a floating gate compatible with CMOS, known as Fishbone-in-Cage Capacitor (FiCC), for its NV component. The evaluations include retention time measurement at various store times and shmoo plots that demonstrate the operational range of the voltage and frequency condition. BET analysis was conducted, accounting for the increased leakage current in addition to store/restore energy. However, the BET formula is mainly based on estimated values derived from SPICE simulations, except for the store current.

3.3 Related work on modeling of MTJ switching characteristics

To estimate the store energy by the TSS control, it is necessary to know how many NVFFs have been successfully stored at the end of the first store operation, and how many have failed and should be stored again. However, the MTJ switching delay time is a stochastic quantity, and analytical models based on complex physical simulations have been proposed.

Zhang et al. derived an analytical model, which assumes that the distribution of the switching delay time resulting from micromagnetic simulations, which can take several days, follows a normal or exponential probability distribution function depending on the magnitude of the store current [40]. Vincent et al. modeled the switching delay time in the intermediate current region, which cannot be represented by the model of Zhang et al. using a gamma distribution [41]. De Rose et al. approximated the results obtained from a combination of micromagnetic and circuit simulations that considered the spatially non-uniform nature of magnetization with a skew-normal probability distribution function [42].

In those previous studies, the analytical models were based on simulation, and no comparison with actual measurement results was made at the system

level. Furthermore, the reliability of the model is unknown if one of the models is selected and applied as it is, because the characteristics of the MTJ implemented on the chip can vary greatly depending on the manufacturer and process.

3.4 Challenges in the previous approaches

As described above, several previously presented NVFFs have not been evaluated in the context of NVPG, and as a result, they have not been quantitatively demonstrated their energy-saving benefits in general intermittent applications. Furthermore, even studies that have conducted breakeven analyses overlooked the increased leakage power caused by additional transistors to control NV elements. Moreover, the characteristics of MTJ implemented in the process may vary greatly between different manufacturers, and it is difficult to guarantee the accuracy of the energy model of VR-NVFF, which takes into account the variability of NVFF in MTJ/CMOS hybrid processes, by directly applying existing simulation models. Therefore, verification based on actual measurements is essential.

The observations described so far motivated us to evaluate a chip implementing VR-NVFF cells to understand the switching characteristics of MTJ in NVFF and to demonstrate the energy reduction effect of the TSS control (in Chapter 4), and then to propose an energy model for NVPG in intermittent operation applications (in Chapter 5), and to finally perform breakeven analysis to obtain guidance to minimize energy consumption of intermittent operation of edge application (in Chapter 6). Note that the modeling construction method used in this study is applicable to other processes, although the analysis is based on actual measurements for the 40 nm MTJ/CMOS hybrid process of Sony Semiconductor Solutions as a case study.

4

Real Chip Measurement and Analysis

VR-NVFF was proposed by Usami et al. [38], and its energy reduction effect was partially evaluated on fabricated chips [60]. The second example of VR-NVFF implementation and the first chip that actually operates at an evaluable level is NVCMA/MC (nonvolatile cool mega array/multicontext) [61]. This chapter presents an overview of NVCMA/MC and investigates the characteristics of VR-NVFF through actual measurements of the implemented chips.

4.1 NVCMA/MC: a chip implementation example of VR-NVFF

4.1.1 Overview of NVCMA/MC

NVCMA/MC is designed as an edge-oriented accelerator with coarse-grained reconfigurable arrays (CGRAs), an reconfigurable architecture that balances power efficiency and flexibility.

(a) CGRA architecture

CGRAs are well-suited architectures for enhancing the performance of compute-intensive applications with limited energy resources by utilizing their coarse-grained reconfigurability and employing hardware acceleration techniques. FPGA,

a type of reconfigurable architecture, offers finer-grained flexibility at the bit-level through the use of look-up tables. However, it requires significant overhead for reconfiguration. On the other hand, CGRAs provide a balanced combination of energy efficiency similar to that of application-specific integrated circuits (ASICs) and moderate flexibility through word-level reconfigurations.

The CGRA architecture on which NVCMA/MC is based is called coolmega array (CMA) architecture [62], which is a straightforward CGRA that forms a straightforward dataflow on the processing element (PE) array (Fig. 4.1) along with PipeRench [63], EGRA [64], DySER [65], and SNAFU [66], among others [67–69].

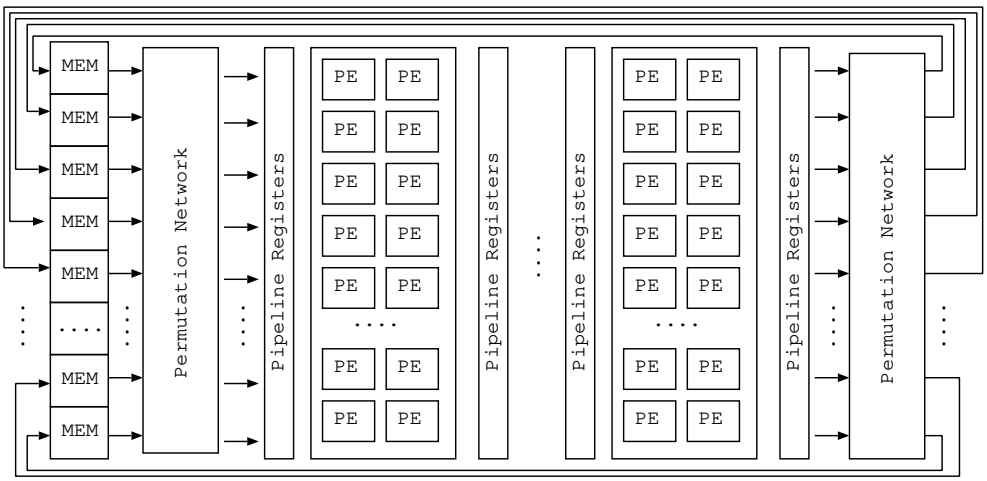


Figure 4.1. Straightforward CGRAs

Fig. 4.2 shows the architecture of NVCMA/MC. A dedicated microcontroller controls the data transfer between the data memory and the PE array with a RISC-type ISA. Configurations for PEs and their interfaces are statically set and reconfigured on a task-by-task basis. Each PE in CMA families does not have a register file, and no clock distribution is needed, thereby reducing dynamic power consumption. Four configuration memory supports the multicontext feature that switches between four different configurations to form a datapath on the PE array for each task, which is similar to DySER and SNAFU in this respect. NVCMA/MC is unique in that it utilizes NVM to retain the configuration data so that the useless leakage power consumption of non-used memories is effectively reduced while other context is in use, which is evaluated in [61].

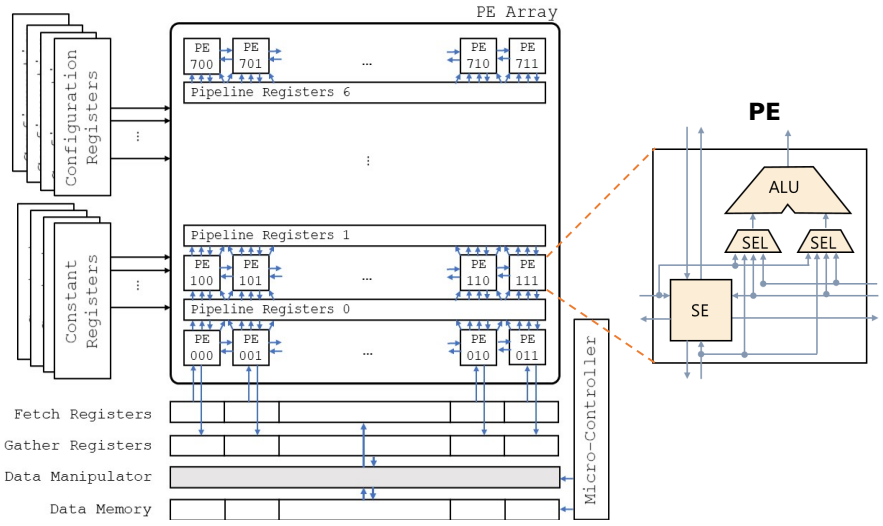


Figure 4.2. NVCMA/MC architectures and its PE.

(b) Introduction of VR-NVFF

In addition to the aforementioned architectural elaboration, NVCMA/MC replaces all memory elements with VR-NVFF to further reduce energy consumption.

NVCMA/MC contains 54,428 NVFFs, which are segmented by two units: power domains (PDs) and store domains (SDs). There are 6 PDs and 24 SDs, as summarized in Table 4.1.

Table 4.1. Power domains and store domains of NVCMA/MC.

Power Domain	Store Domain	Descriptions	# of NVFFs N_{SD}
1	0-1	Instruction memory for μ-controller	4096
2	2-4	Configuration memory for context #0	7111
3	5-7	Configuration memory for context #1	7111
4	8-10	Configuration memory for context #2	7111
5	11-14	Configuration memory for context #3	7111
6	15-18	Data memory #0	10272
	19-23	Data memory #1	11616

PD, a unit of power supply, is independently turned on/off to realize partial PG. NVFFs in the same SD, a unit of controlling NVFF operations, share the control signals to perform the verify, restore, and store operation. The number of NVFFs per SD (N_{SD}) varies from 800 to 4416 in NVCMA/MC. Of the 24 SDs, 9 are allocated to data memories, 2 to instruction memories of the microcontroller, and the remaining 13 to four configuration memories, contexts #0-3. NVFFs for different hardware contexts are mapped to different SD/PD, allowing the store/restore control control to be applied to the configuration data for inactive contexts.

(c) Redundancy of NVCMA/MC

NVCMA/MC adopts two redundancy techniques to avoid malfunctions due to MTJ element failures. One solution is the adoption of an error correction code (ECC) for certain parts of the memory, which allows for 3-bit errors by extending 12 bits to 23 bits using Golay code. The availability was proved to be improved by Ikezoe et al. [70] in exchange for about 2.3 times of hardware overhead for extended bits and the encode/decode circuit. The other solution is the architectural redundancy for the configuration data of the PE array provided by a majority-logic circuit using three multi-configuration data to ensure one reliable configuration data. These redundancy features are beyond the scope of this study, but generally the availability of NV elements and energy efficiency are in a trade-off relationship.

4.1.2 NVPG control with microcontroller

The control mechanism for NVPG in NVCMA/MC is outlined here. As described in Section 2.3.4, a VR-NVFF cell has a total of 10 signal lines (8 NVFF control lines, a PG control line and a clock gating control line). These signal lines in the same SD are controlled all at once by instructions newly extended to the original microcontroller's ISA. Code 4.1 and Code 4.2 are assembly code samples of the NVCMA/MC microcontroller for VR-NVFF control. The instructions in Code 4.1 execute verify and store operations before PG, while those in Code 4.2 execute restore operations after a wake-up from PG. executed before and after PG, respectively. Two important instructions *NVC* and *PG* are described in detail below.

Code 4.1. Assembly code sample of data aware store operation followed by power gating.

```
1  IBM 0 // Select Bitmap Register
2  NVC 0b00101100110 // Start Verify (Stop Clocking)
3  NVC 0b00100100110
4  NVC 0b00000000110 // End Verify
5  NVC 0b00000010110 // Start Store
6  NVC 0b00100010110
7  NVC 0b00000000110 // End Store
8  PGC 0,1,1,1,1,1 // Power Off PD#6
9  DONE
```

Code 4.2. Assembly code sample of restore operation after a wakeup.

```
1  PGC 1,1,1,1,1,1 // Power On PD#6
2  IBM 0 // Select Bitmap Register
3  NVC 0b00110001110 // Start Restore
4  NVC 0b00100001110
5  NVC 0b00000000111 // End Restore (Restart Clocking)
6  DONE
```

(a) *NVC* instruction

The *NVC* (NV control) instruction is tasked with controlling the signal lines of VR-NVFF. Fig. 4.3 shows a schematic diagram of the *NVC* instruction that controls NVFF at the SD-level. A bitmap register of 24 bits, set with *IBM* instruction, specifies the SDs to control, while a 10-bit operand of the *NVC* instruction specifies on/off corresponding to the 10 signal lines of the NVFF. The verify, restore, and store operations are realized by a combination of several *NVC* instructions. An example of an assembly script for a 1-clock store operation after a verify operation is shown in Code 4.1, and Code 4.2 shows an assembly script for a restore operation. The *IBM* (initial bitmap) command is engaged to choose a specific bitmap register from a pre-configured set before the execution of the command. By setting the MSB of the operands of *NVC* instructions to 1, multiple SDs can be operated simultaneously.

(b) *PGC* instruction

Each of six PDs incorporates a certain number of store domains as show in Table 4.1. Fig. 4.4 depicts *PGC* instruction with a 6-bit operand corresponding to each power domain, managing its power states independently. Wake-up procedures are controlled by either the internal instructions or signals from outside the chip. At the end of Code 4.1 and the beginning of Code 4.2, *PGC* instruction is used to turn off and on the power domain #6, respectively.

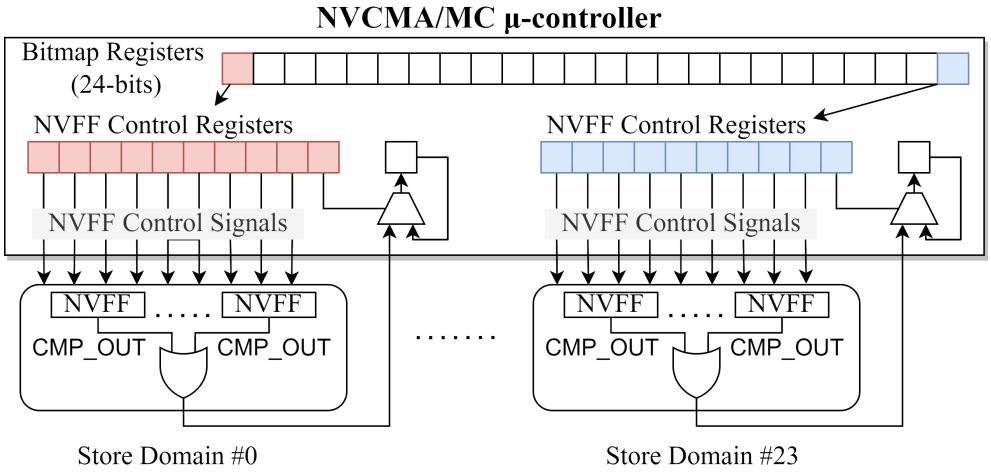


Figure 4.3. Schematic of NVFF control with microcontroller on SD-by-SD basis.

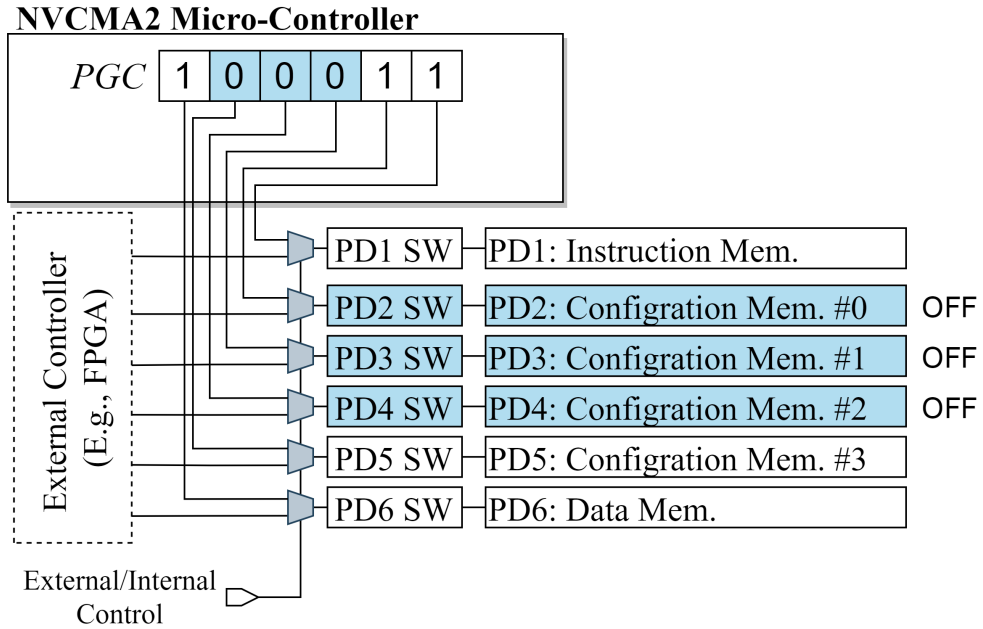


Figure 4.4. NVCMA/MC *PGC* instruction.

4.1.3 Design CAD

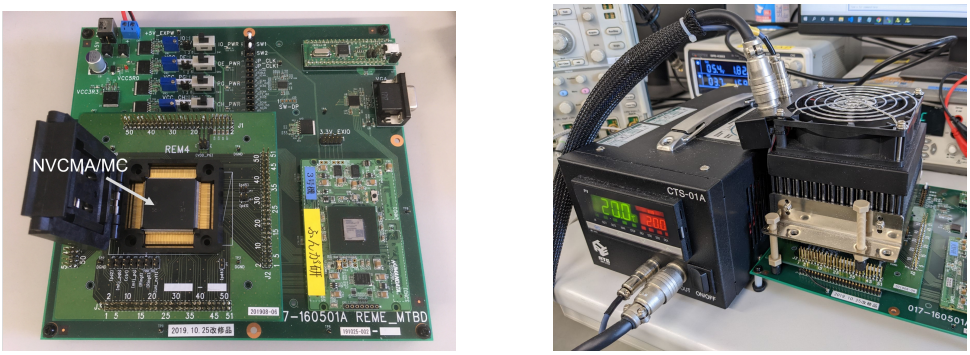
Programming the NVCMA/MC necessitates distinct processes for the PE array and the microcontroller. The process that is executed on the PE array is articulated in C language, and the development environment MENTAI [71] translates it into a dataflow graph. Subsequently, this graph is mapped to the PE array through GenMap [72], a genetic algorithm-based mapping tool. Programmers are presented with various optimal mapping options, each advantageous from different points of view. Meanwhile, the microcontroller's coding, inclusive of NVFF management, is written in assembly code and is translated directly into machine code with dedicated assembly.

4.2 Measurement of implemented VR-NVFF

First, the characteristics of the MTJ switching variability of the implemented VR-NVFF are investigated. Next, the energy reduction effect of the DAS function and the TSS control is demonstrates on an implemented chip.

4.2.1 Fabricated chip and evaluation environment

Fig. 4.5 (a) presents a set of testing and measurement environments, including a fabricated NVCMA/MC chip and the evaluation board. The chip was fabri-



(a) Chip on evaluation board. (b) Peltier thermostat.

Figure 4.5. Evaluation Environment.

cated with a 40 nm perpendicular-MTJ/CMOS hybrid process, and its MTJ size and characteristics are almost identical to those described as "eNVM" in a previous work on MTJs manufactured by Sony Semiconductor Solutions [73].

The fabricated NVCMA/MC chip is placed in the socket of the daughter board and controlled by the host FPGA XCM-208 [74], which shares the motherboard. The motherboard converts the 5V DC power from an external source for the IO and the core of the chip using LDO regulators [75], respectively. In these measurements, to prepare for large current changes during store operations, the daughter board is equipped with additional 2,200 μF electrolytic capacitors for the entire core power, as well as 0.01 μF and 0.022 μF ceramic capacitors and 1,000 μF electrolytic capacitors for each of the six PDs to stabilize the power supply in parallel, other than the default 0.1 μF and 10 μF of electrolytic capacitors.

Switch delay time is known to be significantly and complexly affected by heat [76, 77]. Therefore, the Peltier thermostat presented in Fig. 4.5 (b) is installed to stabilize the measurement environment at 20 ° C, thus eliminating the thermal effect.

Evaluations are performed on a total number of 2,400 NVFFs comprising SD #15, which is a part of the data memory.

4.2.2 Measurement on MTJ switching variations

The MTJ switching variations is evaluated by measuring store pass rates (PR) for various store duration. The PR is defined in Equation (2.6) in Section 2.3.4 (b). Fig. 4.6 plots measured PR at various store duration for V_{DD} = 1.10, 1.15, 1.20 [V], and N_{store} = 600, 1,200, 1,800, and 2,400, describing the MTJ switching variations. The overall trend is that PR in each V_{DD}, N_{store} increases with store duration, indicating that longer store times results in a greater number of MTJs being switched in the store domain.

The results also indicate that PR is positively correlated with V_{DD} and negatively correlated with N_{store}. It stands to reason that PR escalates with an increase in V_{DD} because the current flowing in the MTJ is also higher. On the other hand, the reason why PR worsens with larger N_{store} is possibly attributed to a voltage drop amid the high current demands to store a large number of NVFFs. To investigate the cause of the drop in PR, the oscilloscope waveforms of the transient voltage changes across the core and the target PD are observed at the nearest pins to the chip on the daughter board, respectively, when a store current is applied for 200 ns to 1,200 NVFFs at 1.2 V. Fig. 4.7 (a) shows the default evaluation environment, and Fig. 4.7 (b) shows the observation results in the evaluation environment with additional electrolytic capacitors as described in Section 4.2.1. In both cases, the voltage drops at the moment the store current is applied, and the additional capacitors seem to mitigate the effect to some extent, but not completely eliminate it. The results in Fig. 4.6 are also measured in the state with additional capacitors, and it is reasonable to consider that the current flowing in the MTJ decreases due to the influence of this voltage drop, resulting in deterioration of PR. Although optimization of power supply circuits is beyond the scope of this study, it is safe to say that completely eliminating this effect in resource-constrained edge environments is impossible; therefore, PR should be considered dependent on N_{store}.

Note that there is the upper limit of the operating frequency of the implemented chip (denoted as f_{MAX}) at 80, 83 and 85 [MHz] for V_{DD} = 1.1, 1.15 and 1.20 V, respectively. Therefore, $1/f_{\mathrm{MAX}}$ is the shortest storage duration in each V_{DD}. The operating frequency limitation is most likely due to the I/O part of the chip, not to the VR-NVFF cells.

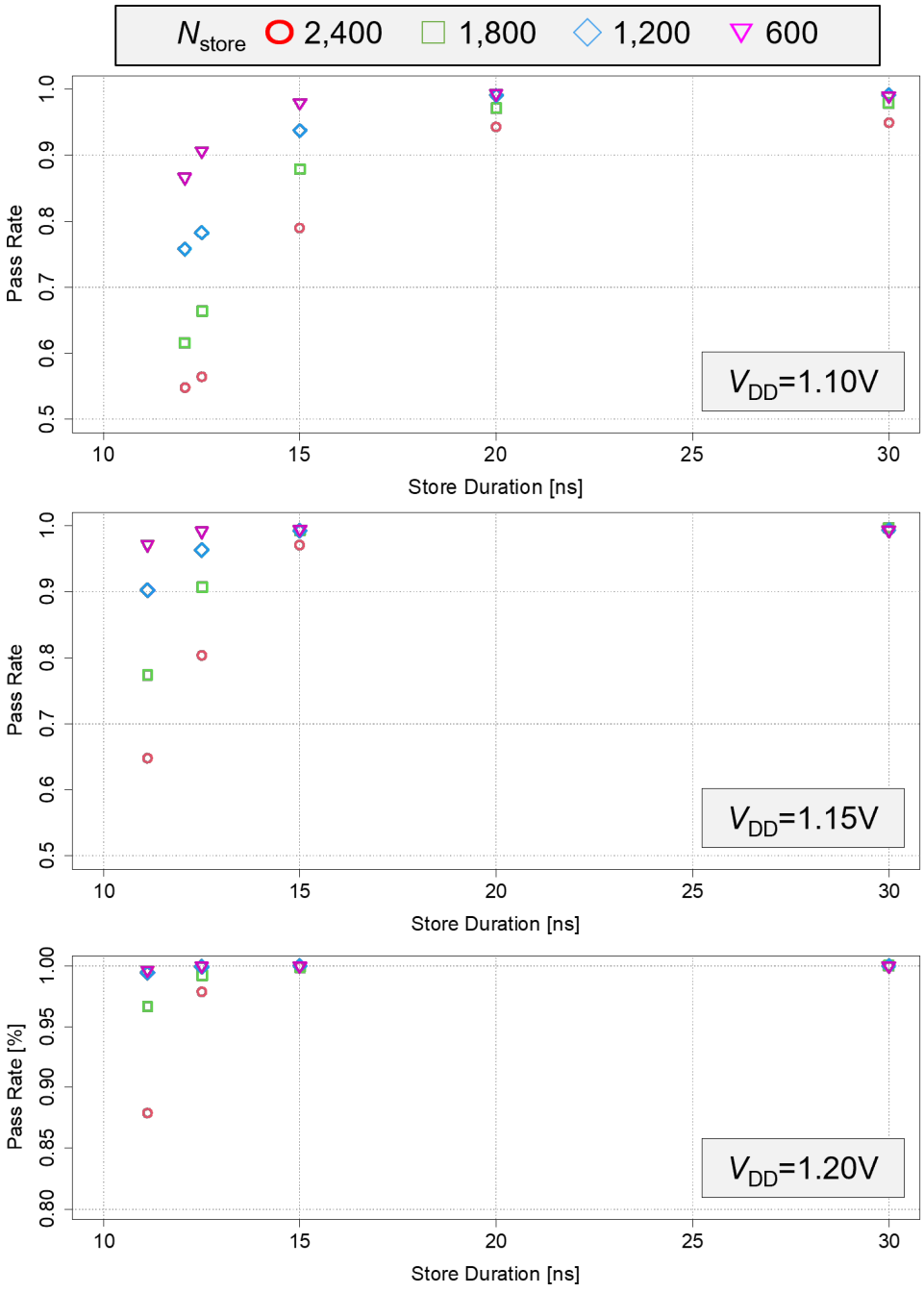
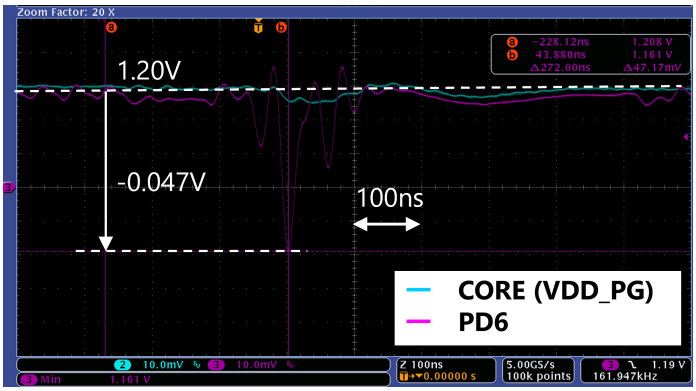
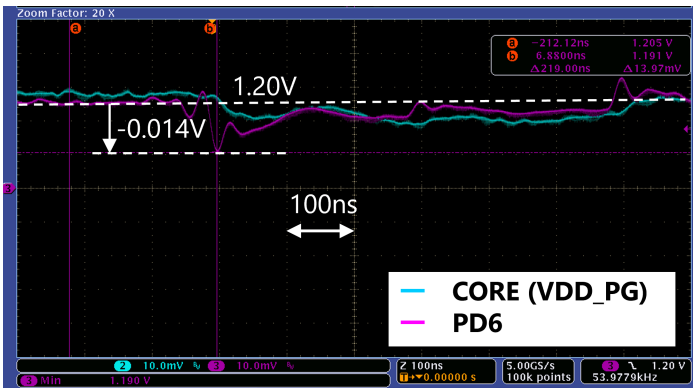


Figure 4.6. Measured pass rate with various V_{DD} and N_{store}.



(a) Measured waveforms in the measurement environment with the daughter board in its default state.



(b) Measured waveforms in the measurement environment where the capacitors are added to the daughter board for power stabilization.

Figure 4.7. Oscilloscope waveforms of the voltage fluctuations of the chip core and PD6 at the instant when stored current is applied to the 1,200 NVFFs in PD6.

4.2.3 Measurement on store energy

Store energy is measured on the implemented chip to demonstrate how much energy reductions are achieved with the DAS function and the TSS control. Let the energy consumed when performing the OSS and the TSS controls be denoted by E_{OSS} and E_{TSS}, respectively.

Although it is true that $E_{\{\mathrm{OSS,TSS}\}} = P_{\{\mathrm{OSS,TSS}\}} \times T_{\{\mathrm{OSS,TSS}\}}$, the time durations for the OSS and the TSS control are extremely short ($<< 1\mu s$), making it impossible to directly measure $P_{\{\mathrm{OSS,TSS}\}}$. Thus, the measurement of storage energy is performed by a calculation using Equation (4.1) with the

average power measured from two different programs with infinite loops as depicted in Fig. 4.8.

$$\begin{aligned} P_{\mathrm{A}} &= \frac{E_{\mathrm{ref}}}{T_{\mathrm{ref}}} \\ P_{\mathrm{B}} &= \frac{E_{\mathrm{ref}} + E_{\{\mathrm{OSS,TSS}\}}}{T_{\mathrm{ref}} + T_{\{\mathrm{OSS,TSS}\}}} \\ \Leftrightarrow E_{\{\mathrm{OSS,TSS}\}} &= P_{\mathrm{B}}(T_{\mathrm{ref}} + T_{\{\mathrm{OSS,TSS}\}}) - P_{\mathrm{A}} T_{\mathrm{ref}} \end{aligned} \tag{4.1}$$

where P_{A} as a reference power is the average power of Program A, which only repeatedly updates data in NVFF latches during T_{ref} while P_{B} is the average power of Program B, which performs store control after the data update operation. E_{OSS} and E_{TSS} include the leakage current of the entire NVCMA/MC chip and the dynamic energy of the microcontroller. The operation part that

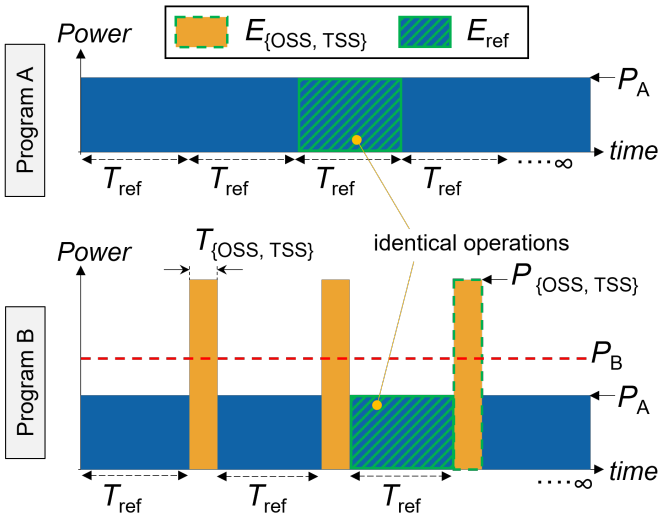


Figure 4.8. Diagram of two programs for actual measurement of store energy

updates the data is exactly the same in Programs A and B. By changing the ratio of data to be updated, $P_{\{\mathrm{OSS,TSS}\}}$ and consequently P_{B} vary in Program B.

Here, the bit update probability, denoted as $BUP = N_{\mathrm{store}}/N_{\mathrm{SD}}$, is defined as a parameter for the data update probability. In general, BUP varies depending on the application and the type of data that the memory handles. For example, kernels of convolutional neural networks do not change unless retrained and updated. On the other hand, feature maps generated from convolution operations frequently update the data in memory, resulting in higher

BUP. As another example, considering a frame buffer for videos, BUP tends to be higher when the input image changes actively, and lower when it is more static, as in the case of surveillance cameras.

Fig. 4.9 and Fig. 4.10 shows measured E_{OSS} and E_{TSS} for $V_{\mathrm{DD}} = 1.10$, 1.20 [V], respectively. T_{req}, enough long time for saturating PR at $V_{\mathrm{DD}} =$ 1.1, 1.15, and, 1.20 V, are assumed as 100, 80, and, 30 [ns], respectively. E_{TSS} was measured at various T_{short} by variously changing the operational clock frequency f [MHz]. T_{long} is set to be T_{req} or above, but slightly varies depending on f.

The red bars in the graph represent E_{OSS}, while the blue bars represent E_{TSS}. The red data labels indicate the energy values when E_{TSS} is minimized by varying T_{short} under each condition. E_{TSS} varies with T_{short}, demonstrating a tendency for E_{TSS} to reach a minimum at an appropriate T_{short} that is neither too long nor too short, as discussed in Section 2.3.4 (c). However, in Fig. 4.9, E_{TSS} does not form complete U-shapes as expected and decreases at T_{short} is 25 ns. The reason we can speculate here is that the operating frequency f of the control signal is relatively high at 80 MHz, leading to a short clock period $1/f$, so as the time for verify, and thus the verify energy is suppressed. The energy breakdown of E_{TSS} cannot be revealed solely by this measurement, but will be investigated by energy modeling later in Chapter 5.

By comparing across different conditions, it is suggested that the optimal T_{short} is shorter with higher V_{DD} or lower BUP. This is because a higher PR is achieved even with a shorter T_{short} under the conditions with less voltage drops, as discussed in Section 4.2.2. This indicates the importance of selecting the appropriate T_{short} according to the conditions. At 1.10 V and $BUP =$ 100 %, E_{TSS} at $T_{\mathrm{short}} = 12.5$ ns is almost double that at the optimal $T_{\mathrm{short}} =$ 17.6 ns, but at BUP is less than 50 %, $T_{\mathrm{short}} = 12.5$ ns seems to be the better choice.

The results also suggest that if V_{DD} is high or BUP is low, the optimal T_{short} would be shorter than the T_{short} used in these measurements. This is because a high PR is more readily achieved even with a shorter store time under these conditions. However, due to the operational frequency constraint of the chip f_{MAX}, this could not be experimentally verified.

E_{OSS} and E_{TSS} and the store energy reduction rate by the TSS control ($\frac{E_{\mathrm{OSS}}-E_{\mathrm{TSS}}}{E_{\mathrm{OSS}}}$) are shown in Fig. 4.11. Each of the E_{TSS} is the minimum E_{TSS} with the optimal T_{short} among Figs. 4.9 and 4.10. The bar graphs of E_{OSS} and E_{TSS} correspond to the left Y-axis, while the line graph of the store energy reduction correlates with the right Y-axis.

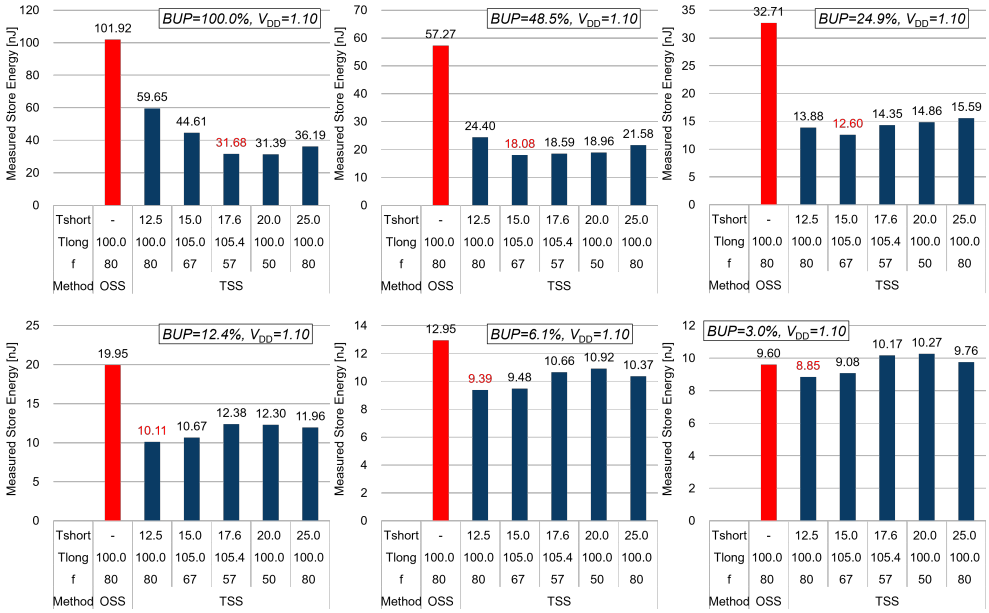


Figure 4.9. Measured store energy E_{OSS} and E_{TSS} at $V_{\mathrm{DD}} = 1.10$ V.

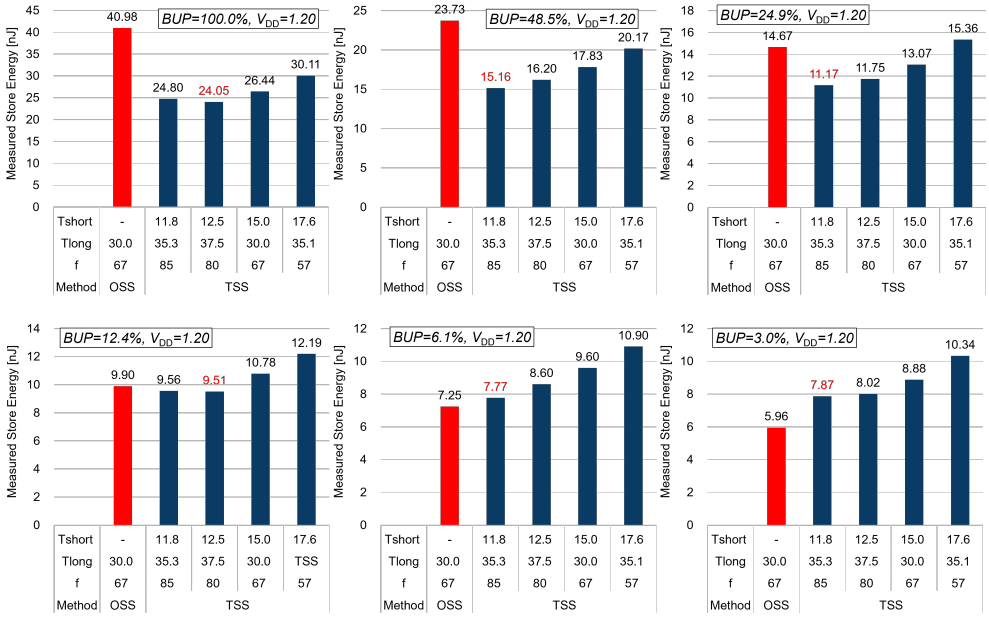
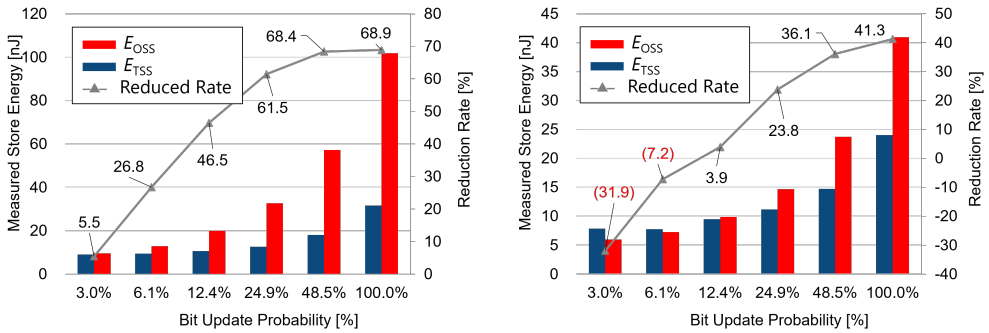


Figure 4.10. Measured store energy E_{OSS} and E_{TSS} at $V_{\mathrm{DD}} = 1.20$ V.



(a) At $V_{DD} = 1.10$.

(b) At $V_{DD} = 1.20$.

Figure 4.11. Measured store energy E_{OSS} and E_{TSS} at optimal T_{short} and store energy reduction rate by the TSS control.

Both E_{OSS} and E_{TSS} decrease with BUP, which is attributed to the DAS function reducing the store energy in NVFFs that do not require store operations. It is also worth mentioning that the reduction rate is more significant when BUP is higher, which means that the number of NVFFs to be stored is large. In contrast, a negative reduction rate is observed, that is, the energy with the TSS control is larger than that with the OSS control, when BUP is small. Although the difference is not as significant, when BUP is approximately less than 10 %, E_{TSS} becomes greater than E_{OSS} at $V_{DD} = 1.20$ V. Again, this is only speculation here, but this inversion phenomenon possibly attributed to the overhead of the TSS control including an additional verify operation, outweighing the energy savings from TSS control. It should also be noted that the inversion does not occur at $V_{DD} = 1.10$ V under the measurement conditions.

The above obsession results suggest that, to minimize energy consumption, it is necessary to estimate whether TSS control can achieve the energy reduction, and how to set the store duration to maximize the energy reduction effect of the TSS control under given conditions specifically in the target application.

4.3 Summary

This chapter presents a detailed examination of the implementation of VR-NVFF using the NVCMA/MC chip as a case study. The chapter focuses on exploring the characteristics of VR-NVFF through actual measurements and analyses of implemented chips.

NVCMA/MC is described as an edge-oriented accelerator, designed with a CGRA architecture, integrating more than 50,000 VR-NVFFs as all memory elements of the chip architecture. The control mechanism for NVPG is detailed, including the concepts of store domains and power domains, and their management methods using the dedicated instructions.

A significant part of the chapter is dedicated to the measurement and analysis of the implemented VR-NVFFs within the NVCMA/MC chip. This includes an examination of the MTJ switching variability and the effects of different operating conditions on the store pass rates and store energy consumption in the TSS control. Importantly, the chapter underscores the necessity of estimating the energy reduction potential of the TSS control in target applications to ensure minimized energy consumption. The actual measurement results from the NVCMA/MC chip demonstrate the importance of such estimations. This emphasizes the need for modeling and analysis to maximize the benefits of VR-NVFF technology in practical applications. These findings contribute significantly to understanding of VR-NVFF technology and even MTJ-based NVFF, also providing valuable insights into system design that incorporates them.

5

Energy Model for Intermittent Operation Applications

In this chapter, we propose an energy model for the estimated energy for the assumed application in the system design phase. On the basis of the measured results of VR-NVFF, the variation of the MTJ switching delay time is modeled and incorporated into the energy model to enable estimation of the store energy of the TSS control. For comparison, energy models for ordinary volatile FF (VFF), SSR-NVFF, and retention FF with a balloon latch (RFF) are also defined, assuming a CMOS process equivalent to that of NVCMA/MC.

5.1 Energy model for VR-NVFF

The energy model of VR-NVFF based on the measurement results is proposed and evaluated in this section. First, the energy model is formulated in Section 5.1.1, and then the model parameters are determined from the measurement results in Section 5.1.3. Finally, the validity of the model is demonstrated by comparing the store energy estimated by the model with the measurement results from Chapter 4.

5.1.1 Formulation of energy model

The total energy consumption of VR-NVFF cells in the unit of store domain in intermittent operation application is formulated. The intermittent operation application as shown in Fig. 5.1 is a repetition of the normal operation period T_{OP} and the idle period T_{NOP}, and the energy model defines the energy E_{cyc} consumed during one cycle (T_{cyc} denoted by Equation (5.1)).

$$T_{\mathrm{cyc}} = T_{\mathrm{OP}} + T_{\mathrm{NOP}} \tag{5.1}$$

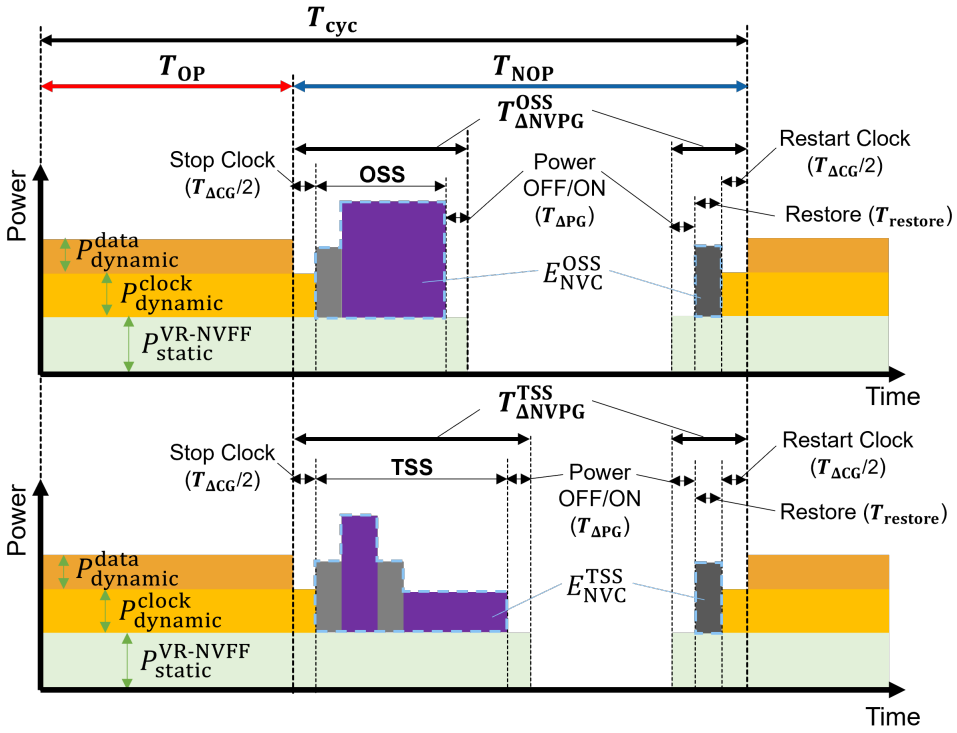


Figure 5.1. Power transition and energy composition in NVPG using VR-NVFF in intermittent operation application.

Here, the operating frequency of the normal operation is f_{OP}, and the operating frequency of the NVPG control during the idle period is f_{NVPG},

and Equation (5.2) holds.

$$\begin{aligned} T_{\mathrm{OP}} &= k \times \frac{1}{f_{\mathrm{OP}}}, \quad k \in \mathbb{Z} \\ T_{\mathrm{NOP}} &= l \times \frac{1}{f_{\mathrm{NVPG}}}, \quad l \in \mathbb{Z} \end{aligned} \tag{5.2}$$

where k and l are arbitrary integers.

For VR-NVFF, the energy consumed in one cycle of intermittent operation is defined as $E_{\mathrm{cyc}}^{\mathrm{OSS}}$ and $E_{\mathrm{cyc}}^{\mathrm{TSS}}$ when OSS and TSS are adopted, respectively, as shown in Equation (5.3).

$$\begin{aligned} E_{\mathrm{cyc}}^{\mathrm{OSS}} &= E_{\mathrm{dynamic}} + E_{\mathrm{static}}^{\mathrm{OSS}} + E_{\mathrm{NVC}}^{\mathrm{OSS}} \\ E_{\mathrm{cyc}}^{\mathrm{TSS}} &= E_{\mathrm{dynamic}} + E_{\mathrm{static}}^{\mathrm{TSS}} + E_{\mathrm{NVC}}^{\mathrm{TSS}} \end{aligned} \tag{5.3}$$

where E_{dynamic} is the dynamic energy defined by Equation (5.4), $E_{\mathrm{static}}^{\mathrm{OSS}}$ and $E_{\mathrm{static}}^{\mathrm{TSS}}$ are the static leakage energies defined by Equation (5.8), and $E_{\mathrm{NVC}}^{\mathrm{OSS}}$ and $E_{\mathrm{NVC}}^{\mathrm{TSS}}$ are the energy for controlling NV elements defined by Equation (5.12).

(a) Dynamic energy

The dynamic power of CMOS consists of a constant power depending on the frequency due to the switching of the clock signal and a power that increases or decreases depending on the switching probability α of the input data to the FF. The first component of E_{cyc}, the dynamic energy, can be expressed as the sum of the two components as shown in Equation (5.4).

$$E_{\mathrm{dynamic}} = E_{\mathrm{dynamic}}^{\mathrm{clock}} + E_{\mathrm{dynamic}}^{\mathrm{data}} \tag{5.4}$$

where $E_{\mathrm{dynamic}}^{\mathrm{clock}}$ and $E_{\mathrm{dynamic}}^{\mathrm{data}}$ are defined by Equation (5.5).

$$\begin{aligned} E_{\mathrm{dynamic}}^{\mathrm{clock}} &= P_{\mathrm{dynamic}}^{\mathrm{clock}} \times (T_{\mathrm{OP}} + T_{\Delta \mathrm{CG}}) \\ E_{\mathrm{dynamic}}^{\mathrm{data}} &= P_{\mathrm{dynamic}}^{\mathrm{data}} \times T_{\mathrm{OP}} \end{aligned} \tag{5.5}$$

where power consumption due to clock switching $P_{\mathrm{dynamic}}^{\mathrm{clock}}$ and power consumption due to input data switching $P_{\mathrm{dynamic}}^{\mathrm{data}}$ are defined by Equation (5.6),

and $T_{\Delta CG}$ is the delay time required for stopping/resuming the clock supply to the FFs with execution of IBM and NVC instructions in the case of NVCMA/MC microcontroller.

$$\begin{aligned} P_{\text{dynamic}}^{\text{clock}} &= I_{\text{dynamic}}^{\text{clock}}(f_{\text{ref}}) \times \frac{f_{\text{OP}}}{f_{\text{ref}}} \times V_{\text{DD}} \times N_{\text{SD}} \\ P_{\text{dynamic}}^{\text{data}} &= \alpha I_{\text{dynamic}}^{\text{data}}(f_{\text{ref}}) \times \frac{f_{\text{OP}}}{f_{\text{ref}}} \times V_{\text{DD}} \times N_{\text{SD}} \end{aligned} \tag{5.6}$$

where N_{SD} is the number of FF cells per store domain, $I_{\text{dynamic}}^{\text{clock}}(f_{\text{ref}})$ is the current consumption per cell due to clock switching measured under the reference operating frequency f_{ref}, and $I_{\text{dynamic}}^{\text{data}}(f_{\text{ref}})$ is the current consumption per cell due to the switching of input data measured under the reference operating frequency f_{ref} and switching activity α. In general, the dynamic current is mainly dominated by the switching current rather than the short-circuit current when operating at a sufficiently higher V_{DD} than the threshold voltage, and each dynamic current is modeled by Equation (5.7).

$$\begin{aligned} I_{\text{dynamic}}^{\text{clock}} &= k_1 V_{\text{DD}} \\ I_{\text{dynamic}}^{\text{data}} &= k_2 \alpha V_{\text{DD}} \end{aligned} \tag{5.7}$$

where k_1 and k_2 are constants. The coefficients of this model are determined from the measurement results in Section 5.1.3.

(b) Static energy

The static energy E_{static} is defined by Equation (5.8) which represents the increase in proportion to the time when the cell is supplied with power.

$$\begin{aligned} E_{\text{static}}^{\text{OSS}} &= P_{\text{static}}^{\text{VR-NVFF}} \times (T_{\text{OP}} + T_{\Delta\text{NVPG}}^{\text{OSS}}) \times N_{\text{SD}} \\ E_{\text{static}}^{\text{TSS}} &= P_{\text{static}}^{\text{VR-NVFF}} \times (T_{\text{OP}} + T_{\Delta\text{NVPG}}^{\text{TSS}}) \times N_{\text{SD}} \end{aligned} \tag{5.8}$$

where $P_{\text{static}}^{\text{VR-NVFF}}$ is the leakage power of the VR-NVFF cell obtained by Equation (5.10), and $T_{\Delta\text{NVPG}}^{\text{OSS}}$ and $T_{\Delta\text{NVPG}}^{\text{TSS}}$ are the time required for NVPG control defined by Equation (5.9) when OSS and TSS are operated, respectively. In the intermittent operation application, the power gating is applied for the time obtained by subtracting the delay time from the idle period of the application

$(T_{\mathrm{NOP}} - T_{\Delta\mathrm{NVPG}}^{\{\mathrm{OSS,TSS}\}})$, and the leakage current becomes zero.

$$\begin{aligned} T_{\Delta\mathrm{NVPG}}^{\mathrm{OSS}} &= T_{\Delta\mathrm{CG}} + T_{\Delta\mathrm{PG}} + T_{\mathrm{verify}} + 2T_{\mathrm{long}} + T_{\mathrm{restore}} \\ T_{\Delta\mathrm{NVPG}}^{\mathrm{TSS}} &= T_{\Delta\mathrm{CG}} + T_{\Delta\mathrm{PG}} + 2T_{\mathrm{verify}} + 2T_{\mathrm{short}} + 2T_{\mathrm{long}} + T_{\mathrm{restore}} \end{aligned} \tag{5.9}$$

where $T_{\Delta\mathrm{PG}}$ is the delay time required for switching the power switch for power gating, and in the case of NVCMA/MC, the *PGC* instruction of the microcontroller is executed as described in Section 5.1.4. T_{verify}, T_{restore}, T_{short}, and T_{long} are the time required to verify, restore, short- and long-store in VR-NVFF, respectively. The reason why 2 is multiplied by T_{short} and T_{long} is that the control mechanism stores the two MTJs in the NVFF cell one by one as depicted in Fig. 2.12 in Chapter 2.

$$P_{\mathrm{static}}^{\text{VR-NVFF}} = I_{\mathrm{static}}^{\text{VR-NVFF}} \times V_{\mathrm{DD}} \tag{5.10}$$

where $I_{\mathrm{static}}^{\text{VR-NVFF}}$ is the leakage current per VR-NVFF cell. The static current is assumed to be modeled by Equation (5.11) for any V_{DD}.

$$I_{\mathrm{static}}^{\text{VR-NVFF}} = k_3 V_{\mathrm{DD}} - B_3 \tag{5.11}$$

where k_3 and B_3 are constants. The coefficients of this model are determined from the measurement results in Section 5.1.3.

(c) NVPG control energy

NVPG control energy, the third component of E_{cyc}, is defined by Equation (5.12).

$$\begin{aligned} E_{\mathrm{NVC}}^{\mathrm{OSS}} &= E_{\mathrm{verify}} + E_{\mathrm{long}}^{\mathrm{OSS}} + E_{\mathrm{restore}} \\ E_{\mathrm{NVC}}^{\mathrm{TSS}} &= 2E_{\mathrm{verify}} + E_{\mathrm{short}}^{\mathrm{TSS}} + E_{\mathrm{long}}^{\mathrm{TSS}} + E_{\mathrm{restore}} \end{aligned} \tag{5.12}$$

where E_{verify} and E_{restore} are the energy consumed by the VR-NVFF cells in the store domain during the verify and restore operations, respectively, and are defined by Equation (5.13). $E_{\mathrm{long}}^{\mathrm{OSS}}$ is the store energy in the OSS control, and $E_{\mathrm{short}}^{\mathrm{TSS}}$ and $E_{\mathrm{short}}^{\mathrm{TSS}}$ are the first and second store energy in the TSS control,

respectively, and are defined by Equation (5.14).

$$\begin{aligned} E_{\text{verify}} &= P_{\text{verify}} \times T_{\text{verify}} \times N_{\text{SD}} \\ E_{\text{restore}} &= P_{\text{restore}} \times T_{\text{restore}} \times N_{\text{SD}} \end{aligned} \tag{5.13}$$

where P_{verify} and P_{restore} are the power consumed by the single NVFF cell in the verify and restore operations, respectively, and are defined by Equation (5.15). N_{SD} is multiplied because verify and restore are performed for all NVFFs in the store domain.

$$\begin{aligned} E_{\text{long}}^{\text{OSS}} &= P_{\text{store}} \times 2T_{\text{long}} \times N_{\text{store}} \\ E_{\text{short}}^{\text{TSS}} &= P_{\text{store}} \times 2T_{\text{short}} \times N_{\text{store}} \\ E_{\text{long}}^{\text{TSS}} &= P_{\text{store}} \times 2T_{\text{long}} \times N_{\text{store}} \times (1 - PR) \end{aligned} \tag{5.14}$$

where P_{store} is the power consumed by the single NVFF cell in the store operation, and is defined by Equation (5.15). In VR-NVFF, the store current flows only in the cells that need to be stored due to the effect of the verify operation, so the number of NVFFs to be stored N_{store} is multiplied. Furthermore, in the second store in the TSS control, the store current flows only in the NVFF cells that failed in the first store, so $N_{\text{store}} \times (1 - PR)$ is multiplied. The estimation of PR will be formulated later in Section 5.1.2.

$$\begin{aligned} P_{\text{verify}} &= I_{\text{verify}} \times V_{\text{DD}} \\ P_{\text{restore}} &= I_{\text{restore}} \times V_{\text{DD}} \\ P_{\text{store}} &= I_{\text{store}} \times V_{\text{DD}} \end{aligned} \tag{5.15}$$

where I_{verify}, I_{restore}, and I_{store} are the current consumed by the single NVFF cell in the verify, restore, and store operations, respectively. Each current is the drain current that flows between the source and drain of the CMOS transistor when the gate is opened. Since I-V characteristics of the drain current can be approximated by a linear function when the transistor is strongly ON [20], we assumed that I_{verify}, I_{restore}, and I_{store} can be modeled using Equation (5.16).

$$\begin{aligned} I_{\text{verify}} &= k_4 V_{\text{DD}} - B_4 \\ I_{\text{store}} &= k_5 V_{\text{DD}} - B_5 \\ I_{\text{restore}} &= k_6 V_{\text{DD}} - B_6 \end{aligned} \tag{5.16}$$

where all k_n and B_n are constants. The coefficients of these models will be determined from the measurement results in Section 5.1.3.

5.1.2 Formulation of MTJ pass rate model

Since the store energy in the TSS control, denoted by Equation (5.12), is the sum of the first and second store energy, it is necessary to accurately estimate PR as the result of the first store.

In Chapter 4, we measured PR under several conditions and found that there is a variation in the switching delay time of the MTJs in the store domain. Thus, the fundamental idea behind PR modeling here is the assumption that the switching delay time of the MTJ can be described by a probability density function (PDF) that belongs to a certain distribution. As a result, PR will be determined by the cumulative distribution function (CDF) of this PDF.

There have been debates regarding the appropriate choice of probability distribution functions to be used to describe the stochastic behavior of the STT switching. Zhang et al. [40,78,79] argues that when the switching current I_{sw} is higher than the critical current I_{c0}, the stochastic nature of the switching delay time of the perpendicular MTJ can be approximated by a normal distribution centered around the mean value $\langle\tau\rangle$ derived from the Landau–Lifshitz–Gilbert (LLG) equations, as described in the formulation provided by Worlegdge et al. [80]. In contrast, the author in the study by Vincent et al. [41] used a gamma distribution, while De Rose et al. [42] suggested alternative models with skew-normal distribution that were better fitted for the intermediate current region, where I_{sw} is relatively close to I_{c0}. The MTJ switching variation comes from the presession of the free-layer magnetization vector. The transient behavior of the MTJ is governed by the torques acting on the vector, and the impact of these torques on magnetization can be simulated using the LLG equation [8]. Of the several torques that act on magnetization, the two dominant torques that affect switching are the spin torque produced by the current flow in the MTJ and the Langevin random field torque produced by the thermal field [81].

Table 5.1 summarizes the main sources of variation that affect torques. Variations in both CMOS and MTJ processes follow a normal distribution [6, 7], and the thermally-induced torque is also assumed to be proportional to the Gaussian noise [8, 9]. Therefore, the variations in both the switching current and the thermal stability are considered to follow a normal distribution. Since all major sources of the variation in switching time are purely random, it is deemed reasonable to approximate it using a Gaussian distribution.

Table 5.1. Primary causes and effects of variations in the MTJ/CMOS hybrid technologies [6–9].

Layer	Causes		Affected Parameters
CMOS	Process Variations	transistor size impurity concentration	threshold voltage drain current
MTJ	Process Variations	surface area tunnel oxide thickness	MTJ resistance switching current
	Process Variations	magnetic anisotropy	switching threshold current density magnetization stability barrier height
	Thermal Fluctuation		Langevin random field torque initial angle of free layer magnetization

Based on the above discussion, this study assumes that PR follows normal distributions with three parameters, including PR_{30}, a value of PR at T_{short} = 30 ns within the practical T_{short} span ($0 < T_{\text{short}}[ns] \leq 30$) as formalized in Equation (5.17) and depicted in Fig. 5.2.

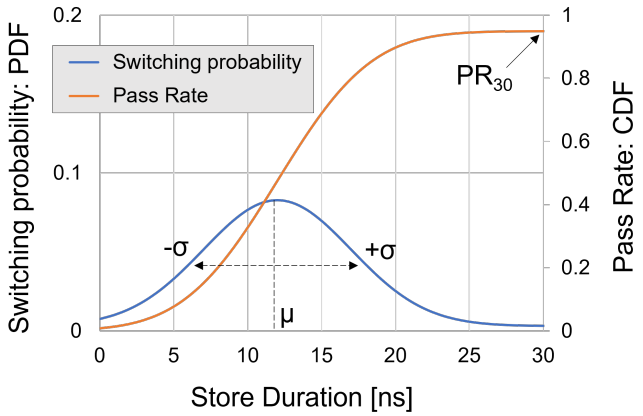


Figure 5.2. An assumed normal distribution model of NVFF's switching delay time. The probability density function (PDF) and the cumulative distribution function (CDF) represent the switching probability and pass rate of a group of NVFFs at a given store duration, respectively.

$$PR(T_{\text{short}}) = \int_0^{T_{\text{short}}} \frac{PR_{30}}{\sqrt{2\pi}\sigma} exp\left(\frac{-(t-\mu)^2}{2\sigma^2}\right) dt \tag{5.17}$$

$$(0 < T_{\text{short}}[ns] \leq 30)$$

Furthermore, as revealed in Chapter 4, PR varies depending on two independent parameters, V_{DD} and N_{store}, so the parameters (μ, σ and PR_{30}) that determine the shape of the PR curve are assumed to be modeled by a multiple regression equation denoted by Equation (5.18).

$$\begin{aligned} \sigma &= \alpha_1 V_{\text{DD}} + \beta_1 N_{\text{store}} + \gamma_1 \\ \mu &= \alpha_2 V_{\text{DD}} + \beta_2 N_{\text{store}} + \gamma_2 \\ PR_{30} &= min(1, \alpha_3 V_{\text{DD}} + \beta_3 N_{\text{store}} + \gamma_3) \end{aligned} \tag{5.18}$$

where the parameters α_n, β_n, and γ_n are determined by multiple regression analysis using the measurement results in Section 5.1.3.

5.1.3 Parameter determination

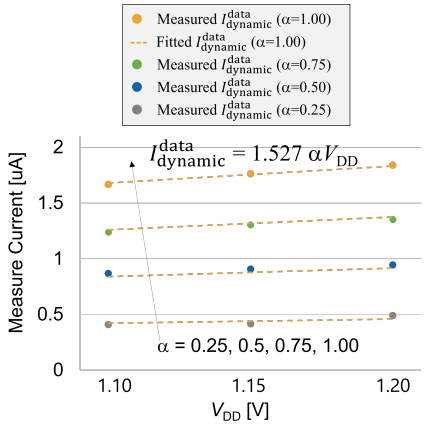
The parameters for current models defined by Equation (5.7), (5.11), and (5.16), and the parameters for pass rate models defined by Equation (5.18) are determined based on the measurement results.

(a) Models for various currents

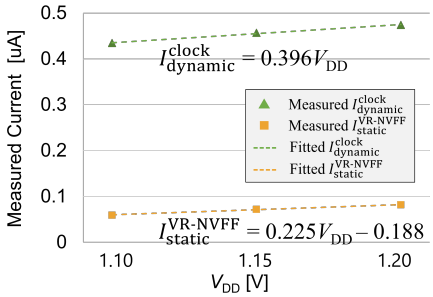
Plots in Fig. 5.3 are the measurement results of $I^{\text{data}}_{\text{dynamic}}$, $I^{\text{clock}}_{\text{dynamic}}$, $I^{\text{VR-NVFF}}_{\text{static}}$, I_{store}, I_{verify} and I_{restore} at different V_{DD}. $I^{\text{data}}_{\text{dynamic}}$ and $I^{\text{clock}}_{\text{dynamic}}$ were measured at frequency $f_{\text{ref}} = 50$ MHz. α varied from 0.25, 0.5, 0.75 and 1.00 for $I^{\text{data}}_{\text{dynamic}}$. The dotted lines are the linear regression approximation lines of the measured currents and are equal to the current models assumed by Equations (5.7), (5.11) and (5.16). The regression lines of $I^{\text{data}}_{\text{dynamic}}$ in Fig. 5.3 (a) are derived from the measurement results at $\alpha = 1$, but they also fit well with the plots for $\alpha = 0.25, 0.5, 0.75$, which is the reasonable property that Equation (2.2) represents. These results support the validity of the current models assumed by linear functions, and the coefficients for each model were determined.

(b) Model for pass rate

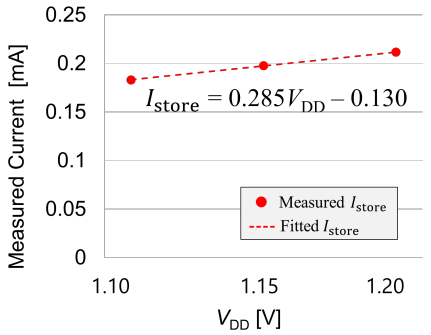
Next, the coefficients of Equation (5.18) are determined based on the measured PR. Plots in Fig. 5.4 are the same data shown in Fig. 4.6 in Chapter 4. The regression curves represented by the solid lines are obtained through the least-squares method assuming that the plots follow the CDF of normal distributions. The strong agreement between the fitted curves and the actual measurement results provides evidence in favor of the PR assumption assuming a Gaussian distribution. The extracted parameters (μ, σ, and PR_{30}) from the fitted lines are summarized as 'fitted parameters' in the inserted tables in Fig. 5.4.



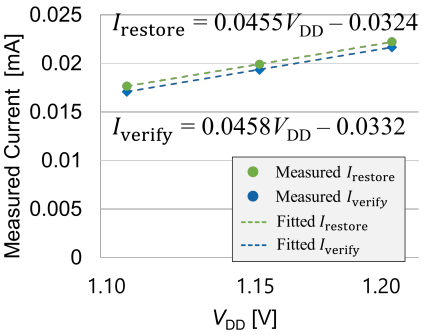
(a) $I^{\text{data}}_{\text{dynamic}}(f_{\text{OP}} = 50$ MHz) with various switching activities α.



(b) $I^{\text{clock}}_{\text{dynamic}}(f_{\text{OP}} = 50$ MHz) and $I^{\text{VR-NVFF}}_{\text{static}}$.



(c) I_{store}.



(d) I_{verify} and I_{restore}.

Figure 5.3. Measured current at each V_{DD} and their linear regression approximation lines.

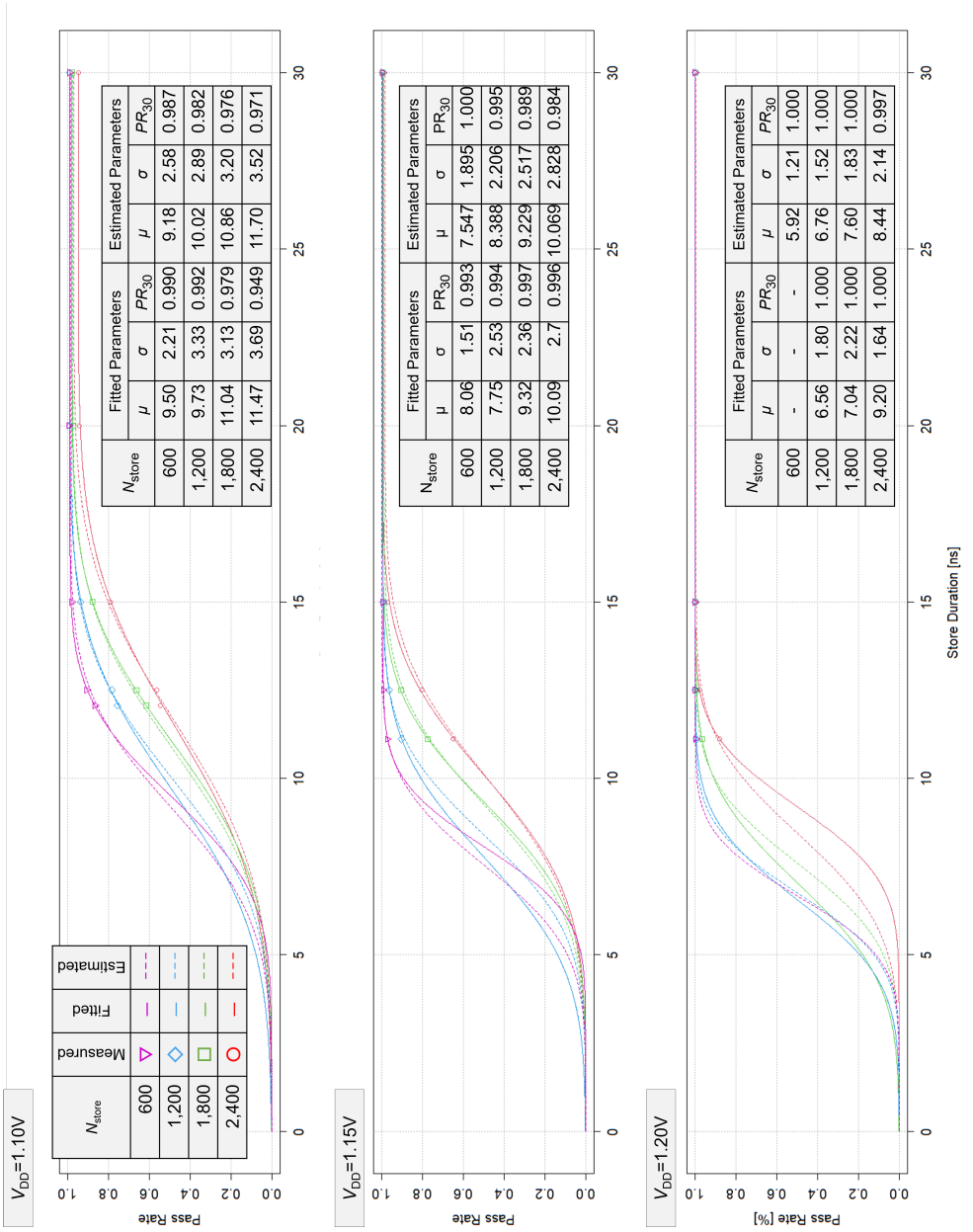


Figure 5.4. Measured pass rates and fitted/estimated pass rates assuming the Gaussian distribution.

Next, multiple regression analyses are performed to determine the coefficients of Equation (5.18). The statistics of the analysis are shown in Table 5.2.

Table 5.2. Results of Multiple regression analysis.

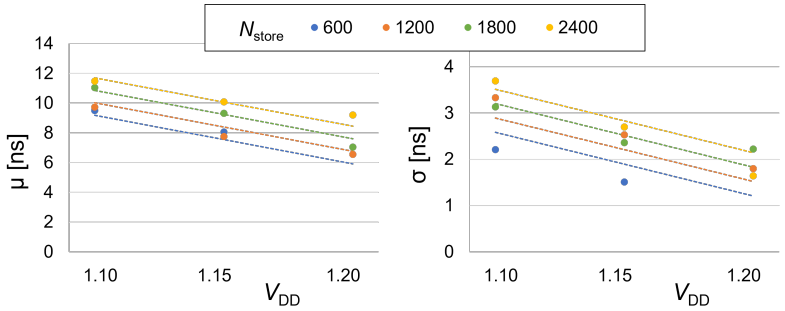
Objective Variable	Adjusted R^2	Significance F	Explanatory Variable	Coefficient	P-value
μ	0.905	0.00003	α_1	-32.564	0.00002
			β_1	0.001	0.00030
			γ_1	44.155	0.00001
σ	0.707	0.00302	α_2	-13.737	0.00159
			β_2	0.001	0.02140
			γ_2	17.382	0.00080
PR_{30}	0.472	0.03180	α_3	25.799	0.01527
			β_3	-0.001	0.11860
			γ_3	70.891	0.00017

The adjusted R^2 is a corrected goodness-of-fit (model accuracy) measure, indicating the model accuracy of the individual parameter prediction models. In this study, we focus more on the significance of each variable than the fitting accuracy of each regression model because the prediction accuracy of PR and store energy is all that matters.

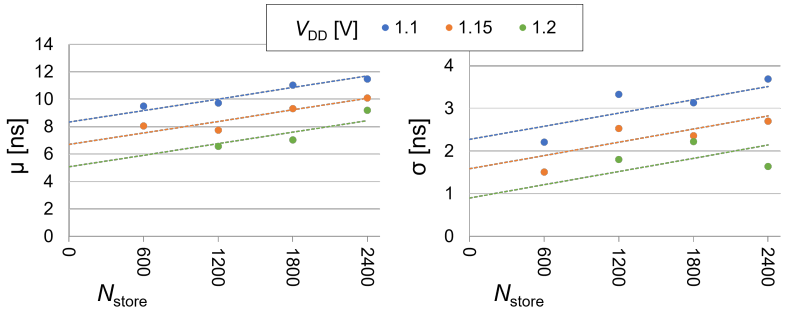
Significance F and P-value indicate the likelihood that a particular outcome occurred randomly. The P-value for each explanatory variable represents the probability that the coefficient is not statically significant. Significance F represents the probability that the combination of the explanatory variables is not statically significant. When the values are less than (typically) 0.05, the result is statistically significant. Here, the results show that all the P values except for β_3 are below 0.05, and all the significance F values are less than 0.05. Therefore, we conclude that the proposed PR model is statistically significant.

In Fig. 5.5, plots are the PR parameters from the measurement results, and the dotted lines are those estimated by the proposed PR model and are shown in the inserted tables as 'Estimated Parameters' in Fig. 5.4 as well. The good fit between them demonstrates the plausibility of the linearity assumption of μ and σ with respect to V_{DD} and N_{store} in the model defined by Equation (5.18) within our evaluation range. Overall, we conclude that the PR parameter estimation by multiple regression analysis with N_{store} and V_{DD} as explanatory

variables is reasonable.



(a) PR parameters varied with V_{DD} at each N_{store}.



(b) PR parameters varied with N_{store} at each V_{DD}.

Figure 5.5. Fitted μ and σ from measured PR and estimated values by the proposed pass rate model.

The PRs estimated the models are described as the dotted curves in Fig. 5.4. In any condition of V_{DD} and N_{store} within the assumed range, the estimated PR fit the measured values well. By using this model, PRs under arbitrary conditions and at any store period can be obtained analytically.

5.1.4 Evaluation: comparison of measured and estimated store energy

The the usefulness of the proposed energy model is evaluated by comparing the estimated energy by the model with the measured energy. The purpose of energy modeling is to help consider the design that minimizes the energy for target applications. To minimize the store energy in VR-NVFF, it is necessary to know which of the OSS control or the TSS control can minimize the energy, and what is the optimal T_{short} in the TSS control. As shown in Chapter 4, it is

a laborious task to measure each condition, and it is not realistic to consider all possibilities. Here, we evaluate whether the proposed energy model can lead to the same conclusion as the measurement results.

Fig. 5.6 and Fig. 5.7 compare the estimated energy and the measured energy for the OSS and TSS control at 1.10 V and 1.20 V, respectively. Since the proposed model is a cell-level model independent of the architecture, the dynamic and static power of NVCMA/MC are subtracted from measured values in this comparison, and only the total energy of verify and store is compared.

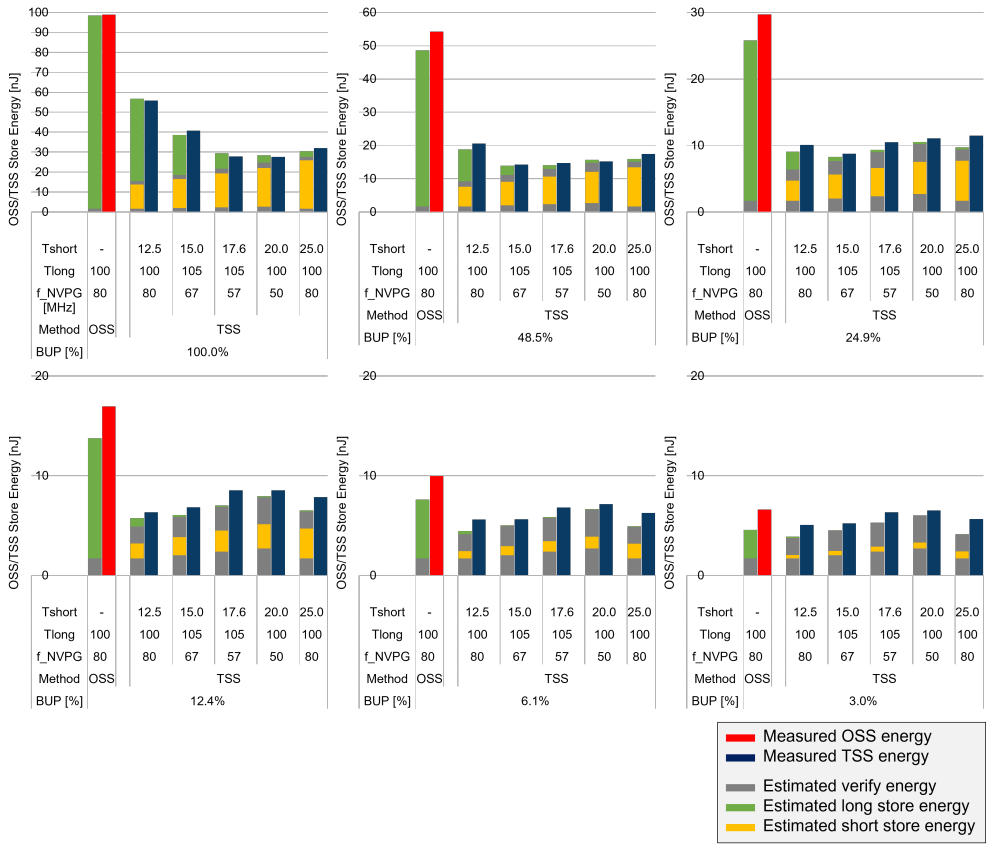


Figure 5.6. Comparison of the estimated store energy with the measurement results for the OSS and the TSS control at $V_{DD} = 1.10$ V.

Red and blue bar charts represent the measured store energy for the OSS and TSS control, respectively. The stacked bar charts represent the estimated energy by the proposed model. The magnitude relations between the store

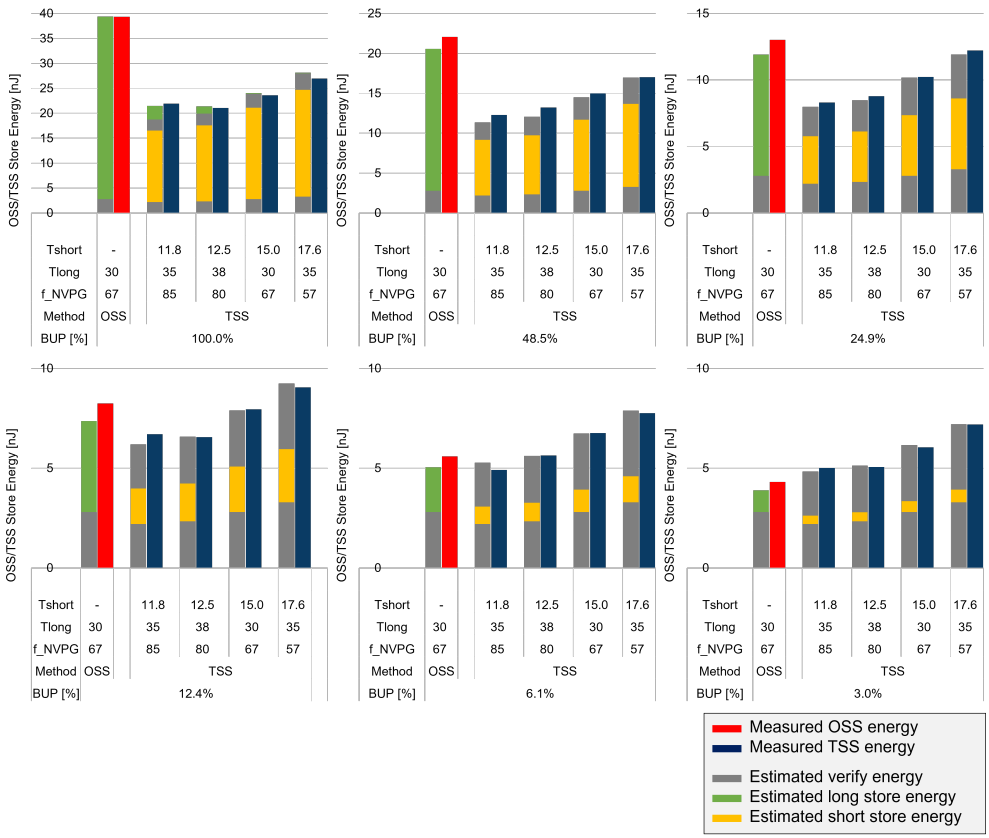


Figure 5.7. Comparison of the estimated store energy with the measurement results for the OSS and the TSS control at $V_{\mathrm{DD}} = 1.20$ V.

energy by the OSS and the TSS control aligns closely between the actual measurements and the model estimation, including the fact that the store energy by the TSS becomes larger when BUP is low at 1.10 V. The estimated energy breakdown clearly confirms the assumption that the overhead of verify operation in the TSS control can be significant, as predicted in Chapter 4. Furthermore, the estimated store energy reflects the effect of the TSS control, which varies with T_{short} as the measured values do. For instance, the optimal T_{short} at 1.20 V and $BUP = 100$ % is determined to be 12.5 ns, while it extends to 20 ns at 1.10 V for the same BUP, which is the same conclusion for both real measurements and model estimates. The estimated results also supports the prediction in Chapter 4 that the energy decreases at $T_{\mathrm{short}} = 25$ ns at 1.10 V are due to the relatively small verify energy at higher operation

frequencies. From the above agreement between the measured and estimated results, it is concluded that the proposed model is useful for selecting which store method to use, or how long T_{short} should be to minimize the store energy under certain conditions.

As outlined in Chapter 4, measuring actual store energy is a complex and time-consuming process, necessitating extra current measurements across two distinct programs for each new condition. In comparison, our developed model enables quick analytical estimations once several key measurements have been taken to determine parameters. Given that real-world applications takes on a various of operation conditions, it is much more practical to use the developed model to simulate the energy consumption of a system running a real application. Thus, we expect our model to be primarily used for identifying the optimal short store time and for aiding in the design of memory systems optimized for variety of applications.

5.1.5 Limitations of the proposed model

As discussed so far, the voltage conditions have been limited to 1.10-1.20 V and the measurement environment has been fixed at 20°C, and therefore the temperature dependence has not been discussed. In this subsection, the voltage range and temperature dependence are discussed to emphasize that the utility of the proposed model is not compromised under these limitations.

(a) Range limit of V_{DD}

Our interest in this paper is to evaluate an actually fabricated chip that incorporates MTJ/CMOS hybrid devices and their peripheral circuits to realize the special storing method. This differentiates our model from many other works that focus on the accurate modeling of simple MTJ devices consisting of a small number of transistors and MTJs. Data based on observations of actual chips are essential to obtain a reliable and practical energy model for such complex systems. Therefore, the evaluation conditions are limited to the operating range of the actually implemented chip, and the energy model based on the measurement results is valid only within the operating range.

We set the upper limit of the voltage at 1.20 V to stay within the rated operational voltage of the chip. The lower limit was set at 1.10 V based on the preliminary evaluation to guarantee MTJ performance with a bit error rate of 1% or less. Fig. 5.8 is a shmoo plot of PR at 20 °C from 1.00 to 1.20

at various store time T_{store}. Below 1.10 V, the PR cannot reach more than 99% even after a considerably long time, which is not a reasonable condition to choose in terms of energy minimization.

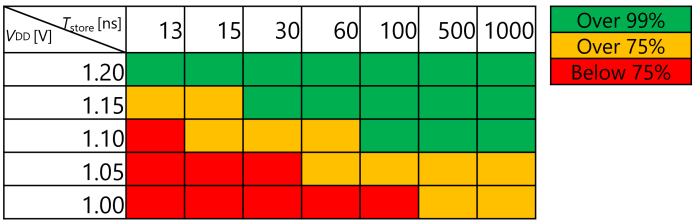


Figure 5.8. Shmoo plot of pass rate from 1.00 to 1.20 V.

The voltage range may seem narrow, but this is the necessary and sufficient range given the purpose of this study because the evaluation conditions are wide enough to cover the range where the implemented chips can operate safely and perform adequately.

(b) Temperature dependency

The temperature dependence of the proposed model is discussed. MTJs are generally known to be temperature dependent, however In MTJ/CMOS hybrid circuits, the effect of temperature on MTJ switching is a combined outcome of two opposing effects. In CMOS circuits, increased temperature causes a decrease in the drain current of transistors, making it harder for switching to happen. On the other hand, regarding the MTJ property, high temperature renders a lower TMR ratio and critical current and higher thermal energy, which all lead to easier switching. Therefore, in the case of considering the model extension regarding temperature, because the balance of the combined effect depends on each implementation and cannot be fully explained by physical theory alone, the model will need to be extended based on measurements, as has been our policy. We can add a new explanatory variable for the temperature to the multiple regression equation for PR (Equation (5.18)) to evaluate and model the effect of temperature on MTJ switching. A logarithmic conversion can be helpful if there are nonlinear effects.

However, we have concluded that the temperature has a trivial effect on the energy estimation in moderate temperature conditions outdoors based on our preliminary evaluation, therefore extending the model regarding temperature is not our priority. Fig. 5.9 shows the results of the preliminary evaluation

on the temperature dependence of the store energy. The bar graphs are the measured store energy at each temperature (-10, 20, 50, 80, and 100°C set by the Peltier thermostat) normalized with E_{conv} at 20°C. The dotted lines indicate the estimated value using our model with parameters determined based on the measurement at 20°C. The bar graphs with stars or red-colored dotted lines indicate the optimal T_{short}s that are measured or estimated to minimize the store energy, respectively. There was no significant difference in

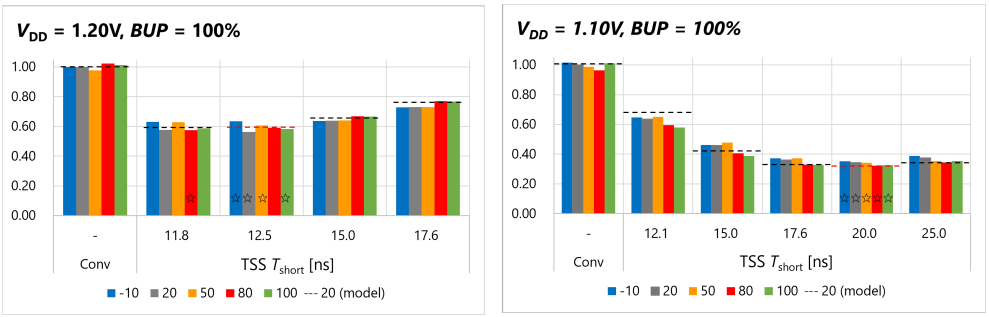


Figure 5.9. Comparison of store energy measured under various temperature conditions and the energy estimated by the model created under 20°C condition.

the measured energy at any temperature compared to the estimated energy from the model created without considering temperature dependence. More precisely, no temperature dependence was identified to the extent that it would affect the selection of the optimal T_{short}. In other words, the temperature dependence of store energy and optimal T_{short} is negligible in the range of -10 to 100 °C, which covers outdoor conditions at most of temperate climate areas in the globe. Therefore, our proposed model, without any further extensions, can satisfy the purpose of choosing the best store time of TSS control and, ultimately, minimizing store energy in the wide temperature range.

5.2 Energy models for alternative FF technologies

In this section, the energy models for various FFs (VFF, RFF, SSR-NVFF) that can be alternative technologies to VR-NVFF in intermittent operation applications. Those technologies have different features as discussed in Chapter 2, thus they may be better choices over VR-NVFF in terms of energy minimization at some conditions. By defining the energy model of alternative FF technologies, quantitative comparison of them can be realized in various conditions. Note that no implemented VFF, RFF, and SSR-NVFF to actually measure are available, following assumptions were made:

1. They are assumed to be implemented in the same process and the same MTJ characteristics as NVCMA/MC (40nm MTJ/CMOS hybrid).
2. Because of that, the power and energy required for MTJ store and restore are the same.
3. The MTJ control part does not assumed to affect the normal operation, therefore the dynamic power consumption is the same among each FF.
4. The static power of each FF is estimated using the ratio between each FF obtained by SPICE simulation and the measured value of VR-NVFF as a reference.
5. The static power of RFF during the idle state is assumed to be 1/5 of that during the active state based on the previous study on RFF [3].

Under above assumption, each energy model is defined by Equation 5.19 for intermittent operation applications as shown in Fig. 5.10. For VFF, RFF, and SSR-NVFF, the energy consumed in one intermittent operation cycle $E_{\mathrm{cyc}}^{\mathrm{VFF}}$, $E_{\mathrm{cyc}}^{\mathrm{RFF}}$, $E_{\mathrm{cyc}}^{\mathrm{SSR}}$ are defined as follows.

$$\begin{aligned} E_{\mathrm{cyc}}^{\mathrm{VFF}} &= E_{\mathrm{dynamic}} + E_{\mathrm{static}}^{\mathrm{VFF}} \\ E_{\mathrm{cyc}}^{\mathrm{RFF}} &= E_{\mathrm{dynamic}} + E_{\mathrm{static}}^{\mathrm{RFF}} + E_{\mathrm{NVC}}^{\mathrm{RFF}} \\ E_{\mathrm{cyc}}^{\mathrm{SSR}} &= E_{\mathrm{dynamic}} + E_{\mathrm{static}}^{\text{SSR-NVFF}} + E_{\mathrm{NVC}}^{\mathrm{SSR}} \end{aligned} \tag{5.19}$$

where E_{dynamic} is the dynamic energy, which is assumed to be common in each FF, including VR-NVFF. $E_{\mathrm{static}}^{\mathrm{VFF}}$, $E_{\mathrm{static}}^{\mathrm{RFF}}$, and $E_{\mathrm{static}}^{\text{SSR-NVFF}}$ are the static energy of VFF, RFF, and SSR-NVFF, respectively, defined by Equation (5.20). $E_{\mathrm{NVC}}^{\mathrm{RFF}}$ and $E_{\mathrm{NVC}}^{\mathrm{SSR}}$ are the energy for controlling NV elements in RFF and SSR-NVFF, respectively, defined by Equation (5.22).

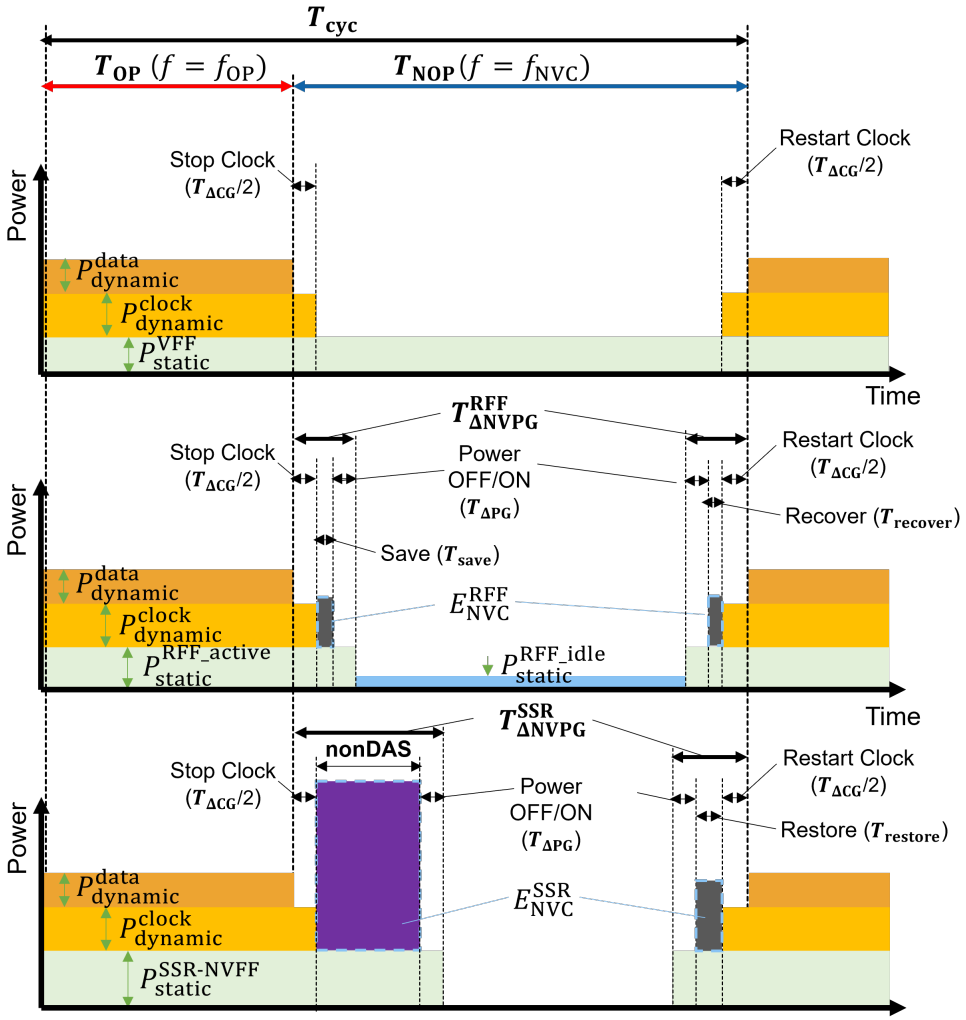


Figure 5.10. Power transition and energy composition in NVPG using alternative FF technologies in intermittent operation application.

$$
\begin{aligned}
E^{\mathrm{VFF}}_{\mathrm{static}} &= P^{\mathrm{VFF}}_{\mathrm{static}} \times (T_{\mathrm{OP}} + T_{\mathrm{NOP}}) \\
E^{\mathrm{RFF}}_{\mathrm{static}} &= P^{\mathrm{RFF\_active}}_{\mathrm{static}} \times (T_{\mathrm{OP}} + T^{\mathrm{RFF}}_{\Delta\mathrm{NVPG}}) + P^{\mathrm{RFF\_idle}}_{\mathrm{static}} \times (T_{\mathrm{NOP}} - T^{\mathrm{RFF}}_{\Delta\mathrm{NVPG}}) \\
E^{\mathrm{SSR\text{-}NVFF}}_{\mathrm{static}} &= P^{\mathrm{SSR\text{-}NVFF}}_{\mathrm{static}} \times (T_{\mathrm{OP}} + T^{\mathrm{SSR}}_{\Delta\mathrm{NVPG}})
\end{aligned}
\tag{5.20}
$$

where $T^{\mathrm{RFF}}_{\Delta\mathrm{NVPG}}$ and $T^{\mathrm{SSR}}_{\Delta\mathrm{NVPG}}$ are the time required for NVPG control, respectively, defined by Equation (5.21). $P^{\mathrm{VFF}}_{\mathrm{static}}$, $P^{\mathrm{RFF\_active}}_{\mathrm{static}}$, and $P^{\mathrm{SSR\text{-}NVFF}}_{\mathrm{static}}$ are the leakage power of VFF, RFF, and SSR-NVFF during the active period, respectively. $P^{\mathrm{RFF\_idle}}_{\mathrm{static}}$ is the leakage power of RFF during idle period where all but a balloon latch in the cell is power gated. Table 5.3 shows the normalized power of each FF with respect to $P^{\mathrm{VR\text{-}NVFF}}_{\mathrm{static}}$, which is extracted from SPICE simulation and assumption from [3]. FreePDK45 [82] provided by NC State University is used for the simulation because the process library used in NVCMA/MC is confidential and not available.

Table 5.3. Normalized simulation results of leakage power for various FF techniques.

Technique	Notation	Normalized leak power
VR-NVFF	$P^{\mathrm{VR\text{-}NVFF}}_{\mathrm{static}}$	1.00
Volatile FF (VFF)	$P^{\mathrm{VFF}}_{\mathrm{static}}$	0.72
Balloon Retention-FF (RFF) (active)	$P^{\mathrm{RFF\_active}}_{\mathrm{static}}$	0.90
Balloon Retention-FF (RFF) (idle)	$P^{\mathrm{RFF\_idle}}_{\mathrm{static}}$	0.18
SSR-NVFF	$P^{\mathrm{SSR\text{-}NVFF}}_{\mathrm{static}}$	0.97

$$\begin{aligned} T^{\mathrm{RFF}}_{\Delta\mathrm{NVPG}} &= T_{\Delta\mathrm{CG}} + T_{\Delta\mathrm{PG}} + T_{\mathrm{save}} + T_{\mathrm{recover}} \\ T^{\mathrm{SSR}}_{\Delta\mathrm{NVPG}} &= T_{\Delta\mathrm{CG}} + T_{\Delta\mathrm{PG}} + 2T_{\mathrm{long}} + T_{\mathrm{restore}} \end{aligned} \tag{5.21}$$

where T_{save} and T_{recover} are the time required for moving date to/from the balloon latch of RFF, respectively, and are assumed to be two clock cycles each.

$$\begin{aligned} E^{\mathrm{RFF}}_{\mathrm{NVC}} &\approx 0 \\ E^{\mathrm{SSR}}_{\mathrm{NVC}} &= E^{\mathrm{SSR}}_{\mathrm{long}} + E_{\mathrm{restore}} \end{aligned} \tag{5.22}$$

where $E^{\mathrm{SSR}}_{\mathrm{long}}$ is the store energy for the non-DAS with SSR-NVFF defined by Equation (5.23). Since NVPG control in RFF does not involve MTJs, therefore $E^{\mathrm{RFF}}_{\mathrm{NVC}} << E^{\mathrm{SSR,OSS,TSS}}_{\mathrm{NVC}}$ holds, we assume that $E^{\mathrm{RFF}}_{\mathrm{NVC}} \approx 0$, which is pessimistic evaluation for MTJ-based NVFF.

$$E_{\text{long}}^{\text{SSR}} = P_{\text{store}} \times 2T_{\text{long}} \times N_{\text{SD}} \tag{5.23}$$

where P_{store} is the power consumed by the single SSR-NVFF cell in the store operation, which is assumed to be the same as that of VR-NVFF since the same MTJ characteristics are assumed.

5.3 Energy estimation using the proposed model

At the end of this chapter, the energy computations of various FFs under various application conditions are estimated using the proposed energy model to learn the characteristics of each FF.

Using the proposed energy model, the energy consumption is estimated under arbitrary operating conditions and intermittent application conditions. Fig. 5.11 is an example of the estimated E_{cyc} of each FF when T_{NOP} and BUP are varied. In VFF and RFF, which consume leakage power during idle time, the energy consumption increases proportionally to T_{NOP}. However, in RFF, the increase is quite slow because the leakage power during idle time is suppressed by the partial power gating. For the three methods with NVFF, there is no dependence on T_{NOP} because the upstream power switch is turned off by NVPG, and the power supply to the NVFF cell is turned off during the application idle period. The difference between VR-NVFF and SSR-NVFF is that VR-NVFF supports data aware store, and the energy is smaller when the BUP is lower in VR-NVFF to save unnecessary store energy, which is not the case in SSR-NVFF. Furthermore, as can be seen by comparing Fig. 5.11 (d) and (e), the store energy is reduced even when the BUP is high as the effect of the TSS control while the OSS control consumes less energy when BUP is close to zero.

From the above discussion, it is clearly suggested that the dependence of the energy consumption of each FF on BUP, T_{NOP} and various conditions is different, and which FF technology minimizes the energy consumption depends on the conditions. Therefore, the system designer must analyze the application conditions and select the best FF technology for the system to realize the energy efficient system.

Note that the energy model is the energy consumption in the NVFF cell in the store domain, which is a set of NVFF cells, as defined in Chapter 5, and does not include the energy consumption in the power switch provided for each NVPG control circuit and the upstream power domain. Since there

are many possibilities depending on the implementation method and the scale of the power domain, this study focuses only on the NVFF cell, and as in previous studies [4, 53] that compare NVFF and RFF, we assume that the energy consumption of NVFF in the PG is zero.

5.4 Summary

In this chapter, we developed an energy model to estimate the energy of NVPG in intermittent operation applications. The model is based on the measured results of VR-NVFF and incorporates the variability in the MTJ switching delay time. For comparative purposes, models for other FF technologies such as volatile FF (VFF), SSR-NVFF, and retention FF with balloon latch (RFF), all assumed to be implemented in a 40nm MTJ/CMOS hybrid process like NVCMA/MC, were also established.

The core achievement of this chapter is the formulation and validation of the energy model for VR-NVFF. Firstly, the models were assumed, and the parameters were determined by empirical data. The VR-NVFF energy model, based on actual measurements, estimates the store energy for both one-step store (OSS) and two-step store (TSS) controls. It incorporates dynamic, static, and NVPG control energies in the NVPG scenario. The linear regression of the measured currents affirmed the validity of the assumed linear current models, and multiple regression analysis was employed to establish the coefficients for the pass rate equation.

A significant contribution is the establishment of a pass rate model for the MTJ, assuming a normal distribution for the variability in switching delay time. This model enables the accurate estimation of the first store's pass rate, crucial for the TSS control's energy estimation. The model is proven to be an efficient tool for selecting the optimal store method and store duration under varied conditions by comparing its estimates and measured results.

In addition to VR-NVFF, energy models for alternative FF technologies like VFF, RFF, and SSR-NVFF are defined, allowing quantitative comparison under various application conditions. These models consider dynamic and static energies and energy for controlling NV elements, with the assumption that all FFs share the same MTJ characteristics as NVCMA/MC. The estimated energy consumption in different NVPG scenarios demonstrates the distinct energy consumption patterns of each FF technology. The results highlight that the choice of the most energy-efficient FF technology varies on the basis of specific application conditions.

In conclusion, the energy model proposed in this chapter is a robust tool for estimating and minimizing store energy in intermittent operation appli-

cations. This model is instrumental for system designers in determining the most energy-efficient memory technologies and control methods, particularly in the context of intermittent operation applications.

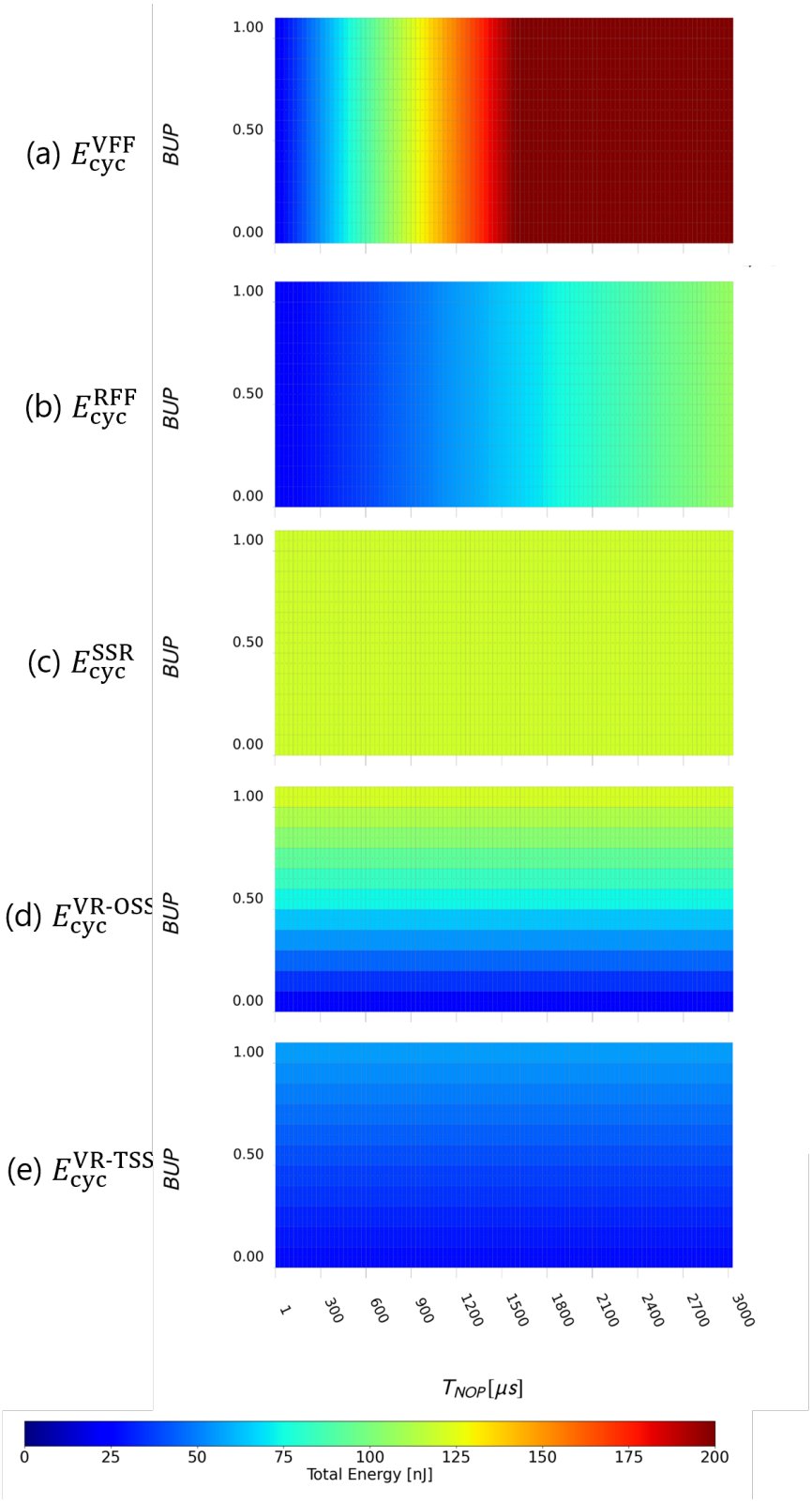


Figure 5.11. Estimated energy E_{cyc} of each FF using the proposed model (V_{DD} = 1.10 V, T_{OP} = 10 μs, f_{OP} = 20 MHz, N_{SD} = 2400, α = 0.2).

6

Breakeven Analysis for Energy Minimization in NVPG

This chapter conducts a breakeven analysis between several FF technologies using the proposed energy model. Breakeven analysis is used to compare profits and overhead between alternative technologies to help system designers make better decisions for minimizing energy consumption. Here, from the options in the previous chapter, we propose an analytical scheme to choose the optimal option that minimizes energy consumption using the proposed model. Breakeven analysis assuming dynamic voltage scaling (DVS), which was not implemented in the NVCMA/MC system, is also briefly discussed using the proposed model.

6.1 Definitions of breakeven point indicators

The breakeven point here refers to the condition where the energy savings and overhead due to NVPG and the TSS, are offset. In this study, T_{OP}, T_{NOP}, and BUP are employed as indicators of the breakeven point. These are metrics that define intermittent operation applications and are naturally determined at the stage of selecting the target application at the system design phase.

In each application, whether these three indicators are smaller or larger than Breakeven T_{OP} (BET_{OP}), Breakeven T_{NOP} (BET_{NOP}), and Breakeven BUP ($BEBUP$) becomes the criterion for deciding which FF to choose.

6.1.1 Breakeven T_{NOP}

BET_{NOP} is an indicator often used to evaluate the performance of energy reduction techniques. It is the breakeven point for intermittent operation applications, where if the non-operating period T_{NOP} is longer than BET_{NOP}, then energy savings are achieved.

Here, $BET_{\mathrm{NOP}}^{\mathrm{X\text{-}Y}}$ denotes BET_{NOP} in the comparison of two options X and Y and is defined as follows.

$$\begin{aligned} T_{\mathrm{NOP}} = BET_{\mathrm{NOP}}^{\mathrm{X\text{-}Y}} &\implies E_{\mathrm{cyc}}^{\mathrm{X}} = E_{\mathrm{cyc}}^{\mathrm{Y}}, \\ T_{\mathrm{NOP}} < BET_{\mathrm{NOP}}^{\mathrm{X\text{-}Y}} &\implies E_{\mathrm{cyc}}^{\mathrm{X}} < E_{\mathrm{cyc}}^{\mathrm{Y}}, \\ T_{\mathrm{NOP}} > BET_{\mathrm{NOP}}^{\mathrm{X\text{-}Y}} &\implies E_{\mathrm{cyc}}^{\mathrm{X}} > E_{\mathrm{cyc}}^{\mathrm{Y}}. \end{aligned} \tag{6.1}$$

Fig. 6.1 compares the power consumption and cumulative energy of two options, one without PG (X) and one with NVPG (Y). It shows the case where $T_{\mathrm{NOP}} = BET_{\mathrm{NOP}}^{\mathrm{X\text{-}Y}}$. The energy consumption of both options is equal (breakeven) in one cycle of the application.

Furthermore, based on the definitions of the energy model, Equations (5.3) and (5.19), $BET_{\mathrm{NOP}}^{\mathrm{X\text{-}Y}}$ can be expressed as follows by breaking it down to two components: 1) the overhead due to the leak current during normal operation ($BET_{\mathrm{leak}}^{\mathrm{X\text{-}Y}}$) and 2) the overhead of NVPG control ($BET_{\mathrm{NVPG}}^{\mathrm{X\text{-}Y}}$).

$$\begin{aligned} BET_{\mathrm{NOP}}^{\mathrm{X\text{-}Y}} &= BET_{\mathrm{leak}}^{\mathrm{X\text{-}Y}} + BET_{\mathrm{NVPG}}^{\mathrm{X\text{-}Y}} \\ &= \eta_{\mathrm{leak}}^{\mathrm{X\text{-}Y}} T_{\mathrm{OP}} + BET_{\mathrm{NVPG}}^{\mathrm{X\text{-}Y}} \end{aligned} \tag{6.2}$$

where $\eta_{\mathrm{leak}}^{\mathrm{X\text{-}Y}}$ represents the relative difference of the static current between X and Y.

When VFF is the baseline X, $BET_{\mathrm{NOP}}^{\mathrm{VFF\text{-}Y}}$, where Y $\in$ {RFF, SSR, OSS,

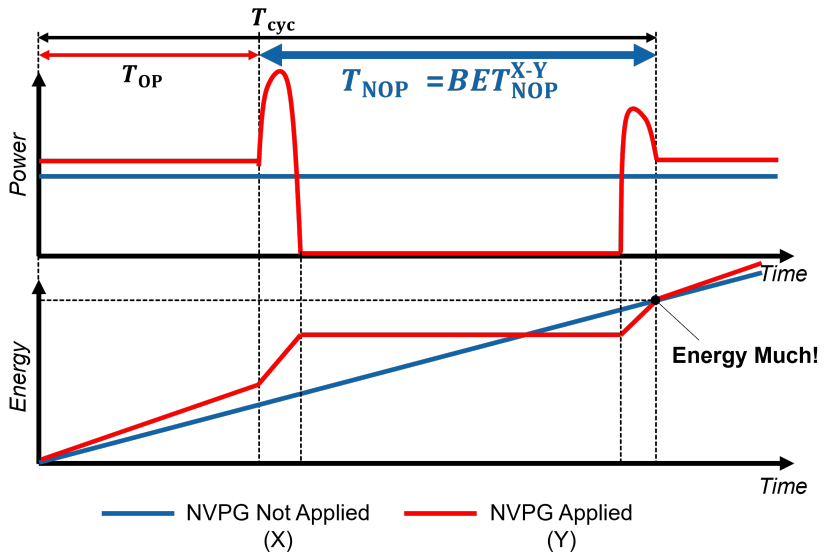


Figure 6.1. Power and energy comparison of two different FF technologies with and without NVPG at breakeven T_{NOP}.

TSS}, is derived from Equation (5.3) as follows.

$$\begin{aligned}
&\text{When } X = \text{VFF}: \\
&\qquad BET^{\text{VFF-Y}}_{\text{NOP}} = \eta^{\text{VFF-Y}}_{\text{leak}} T_{\text{OP}} + BET^{\text{VFF-Y}}_{\text{NVPG}} \\
&\qquad \eta^{\text{VFF-Y}}_{\text{leak}} = \frac{P^{\text{Y}}_{\text{static}} - P^{\text{VFF}}_{\text{static}}}{P^{\text{VFF}}_{\text{static}}} \\
&\qquad BET^{\text{VFF-Y}}_{\text{NVPG}} = \frac{P^{\text{Y}}_{\text{static}} T^{\text{Y}}_{\Delta\text{NVPG}} + E^{\text{Y}}_{\text{NVC}}}{P^{\text{VFF}}_{\text{static}}}
\end{aligned} \tag{6.3}$$

When RFF is the baseline X, $BET^{\text{RFF-Y}}_{\text{NOP}}$, where Y $\in$ {SSR, OSS, TSS}, is derived from Equation (5.19) as follows.

$$\begin{aligned}
&\text{When } X = \text{RFF}: \\
&\qquad \eta^{\text{RFF-Y}}_{\text{leak}} = \frac{P^{\text{Y}}_{\text{static}} - P^{\text{RFF_active}}_{\text{static}}}{P^{\text{RFF_idle}}_{\text{static}}} \\
&\qquad BET^{\text{RFF-Y}}_{\text{NVPG}} = \frac{P^{\text{Y}}_{\text{static}} T^{\text{Y}}_{\Delta\text{NVPG}} + E^{\text{Y}}_{\text{NVC}} - (P^{\text{RFF_active}}_{\text{static}} - P^{\text{RFF_idle}}_{\text{static}}) T^{\text{RFF}}_{\Delta\text{NVPG}}}{P^{\text{RFF_idle}}_{\text{static}}}
\end{aligned} \tag{6.4}$$

6.1.2 Breakeven T_{OP}

When the active leakage power is higher, the longer T_{OP} is, the greater the leakage energy overhead becomes. BET_{OP} is referenced when comparing two FFs with different active leakage power. When the normal operation time T_{OP} of an intermittent operation application is longer than BET_{OP}, the leakage energy overhead of NVFF cells with relatively complex structures and therefore a large area exceeds the energy reduction benefit.

In this study, $BET_{\mathrm{OP}}^{\text{OSS-SSR}}$ and $BET_{\mathrm{OP}}^{\text{TSS-SSR}}$ are used to conduct a breakeven analysis between SSR-NVFF and the OSS with VR-NVFF and the TSS with VR-NVFF, respectively. Those are defined as follows.

$$\begin{aligned} T_{\mathrm{OP}} = BET_{\mathrm{OP}}^{\{\text{OSS,TSS}\}\text{-SSR}} &\Longrightarrow E_{\mathrm{cyc}}^{\{\text{OSS,TSS}\}} = E_{\mathrm{cyc}}^{\mathrm{SSR}}, \\ T_{\mathrm{OP}} < BET_{\mathrm{OP}}^{\{\text{OSS,TSS}\}\text{-SSR}} &\Longrightarrow E_{\mathrm{cyc}}^{\{\text{OSS,TSS}\}} < E_{\mathrm{cyc}}^{\mathrm{SSR}}, \\ T_{\mathrm{OP}} > BET_{\mathrm{OP}}^{\{\text{OSS,TSS}\}\text{-SSR}} &\Longrightarrow E_{\mathrm{cyc}}^{\{\text{OSS,TSS}\}} > E_{\mathrm{cyc}}^{\mathrm{SSR}}. \end{aligned} \tag{6.5}$$

Fig. 6.2 compares the power consumption and cumulative energy of two options, OSS or TSS control with VR-NVFF and non-DAS with SSR-NVFF. It shows the case where $T_{\mathrm{OP}} = BET_{\mathrm{OP}}^{\{\text{OSS,TSS}\}\text{-SSR}}$. The energy consumption of both options is equal (breakeven) in one cycle of the application. When

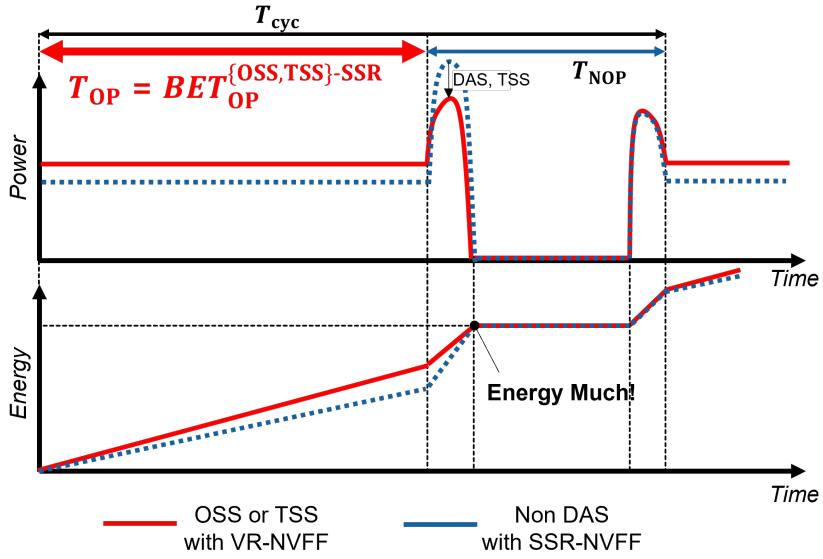


Figure 6.2. Power and energy comparison for VR-NVFF with DAS/TSS control and SSR-NVFF without DAS at breakeven T_{OP}.

$T_{\mathrm{OP}} < BET_{\mathrm{OP}}$, energy saving is achieved with DAS function of VR-NVFF. However, once T_{OP} exceeds BET_{OP}, the increasing energy overhead due to the leak proportional to T_{OP} becomes significant. In that case, adopting SSR-NVFF is a better choice to minimize energy due to its simpler cell structure.

6.1.3 Breakeven BUP

Breakeven BUP ($BEBUP$) is the decision branching point for selecting between two DAS methods, based on BUP for intermittent operation applications. As we learned from the measured results, the energy reduction effects of the OSS control and the TSS control are reversed at certain BUP, and the boundary BUP is defined as $BEBUP^{\text{OSS-TSS}}$ as follows.

$$\begin{aligned} BUP = BEBUP^{\text{OSS-TSS}} &\implies E_{\mathrm{cyc}}^{\mathrm{OSS}} = E_{\mathrm{cyc}}^{\mathrm{TSS}}, \\ BUP < BEBUP^{\text{OSS-TSS}} &\implies E_{\mathrm{cyc}}^{\mathrm{OSS}} < E_{\mathrm{cyc}}^{\mathrm{TSS}}, \\ BUP > BEBUP^{\text{OSS-TSS}} &\implies E_{\mathrm{cyc}}^{\mathrm{OSS}} > E_{\mathrm{cyc}}^{\mathrm{TSS}}. \end{aligned} \tag{6.6}$$

When BUP exceeds $BEBUP^{\text{OSS-TSS}}$, the energy savings from the energy reduction by the TSS outweigh the overhead of the TSS, making the TSS the optimal choice to minimize energy consumption. It should be noted that both the OSS and the TSS are performed with VR-NVFF and can be switched between software-wise. Therefore, the decision on which to choose does not necessarily need to be made in the system design stage.

6.2 Breakeven analysis

This section demonstrates the breakeven analysis, the scheme for selecting energy minimizing techniques using the breakeven criterion introduced above.

First, a breakeven analysis for VFF, SSR-NVFF, and VR-NVFF is performed. Subsequently, the analysis that includes RFF is also carried out. In addition, the energy savings effect of employing DVS in MTJ-based NVFF is also briefly examined.

6.2.1 MTJ-based NVFF vs. VFF

(a) Analysis method

The flowchart in Fig. 6.3 demonstrates how to choose the most energy-efficient option among four choices, 1) non NVPG with VFF (VFF), 2) non DAS with SSR-NVFF (SSR), 3) OSS control with VR-NVFF (OSS), and 4) TSS control with VR-NVFF (TSS). When conditions of target applications (T_{NOP}, T_{OP}, BUP, V_{DD}) are given, three each breakeven point is calculated using the proposed energy model. Simultaneously, the flowchart guides to a unique optimal choice.

The decision making maps illustrated in Fig. 6.4 is almost equivalent to the flowchart in Fig. 6.3, and visually represents the breakeven boundaries of each FF alternatives. Once BUP and V_{DD} of the target application are given, the OSS or the TSS is eliminated from the options using $BEBUP^{\mathrm{OSS\text{-}TSS}}$, and the lines of BET_{NOP} and BET_{OP} can be drawn on the graph with T_{NOP} on the X axis and T_{OP} on the Y axis. BET_{NOP} ($= \eta_{\mathrm{leak}} T_{\mathrm{OP}} + BET_{\mathrm{NVPG}}$) is a linear function with a positive slope η_{leak} and intercept BET_{NVPG}. The slope η_{leak} is determined by the ratio of leakage power between the two FFs, and remains constant regardless of other conditions. Next, any T_{OP} and T_{NOP} are given, one optimal choice that minimizes energy can be determined depending on where it fits in the condition region divided by the breakeven lines.

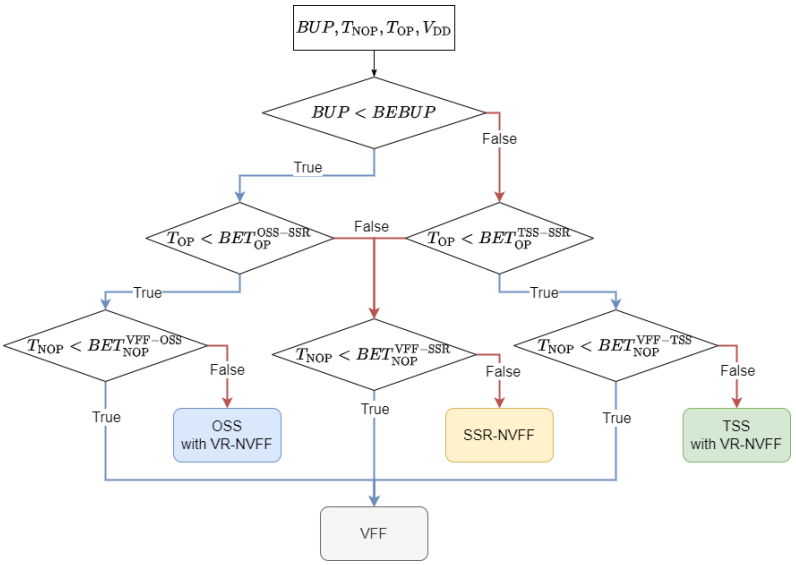


Figure 6.3. The decision flow of breakeven analysis comparing VFF, SSR-NVFF, and VR-NVFF.

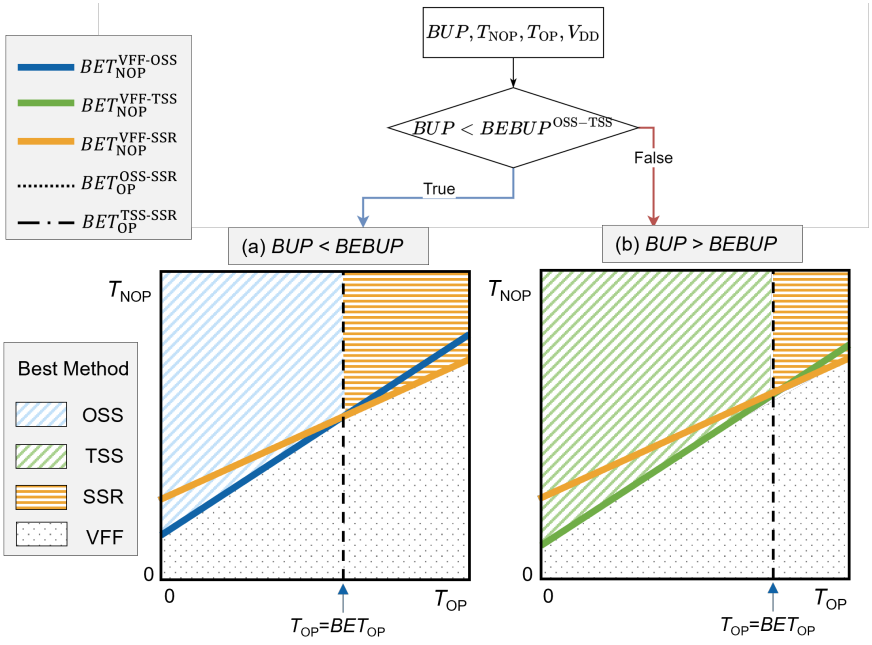


Figure 6.4. The decision making map of breakeven analysis comparing VFF, SSR-NVFF, and VR-NVFF.

(b) Analysis results

The breakeven analysis for comparing VFF, SSR-NVFF, and VR-NVFF with log-log graphs for various BUP and V_{DD} is conducted.

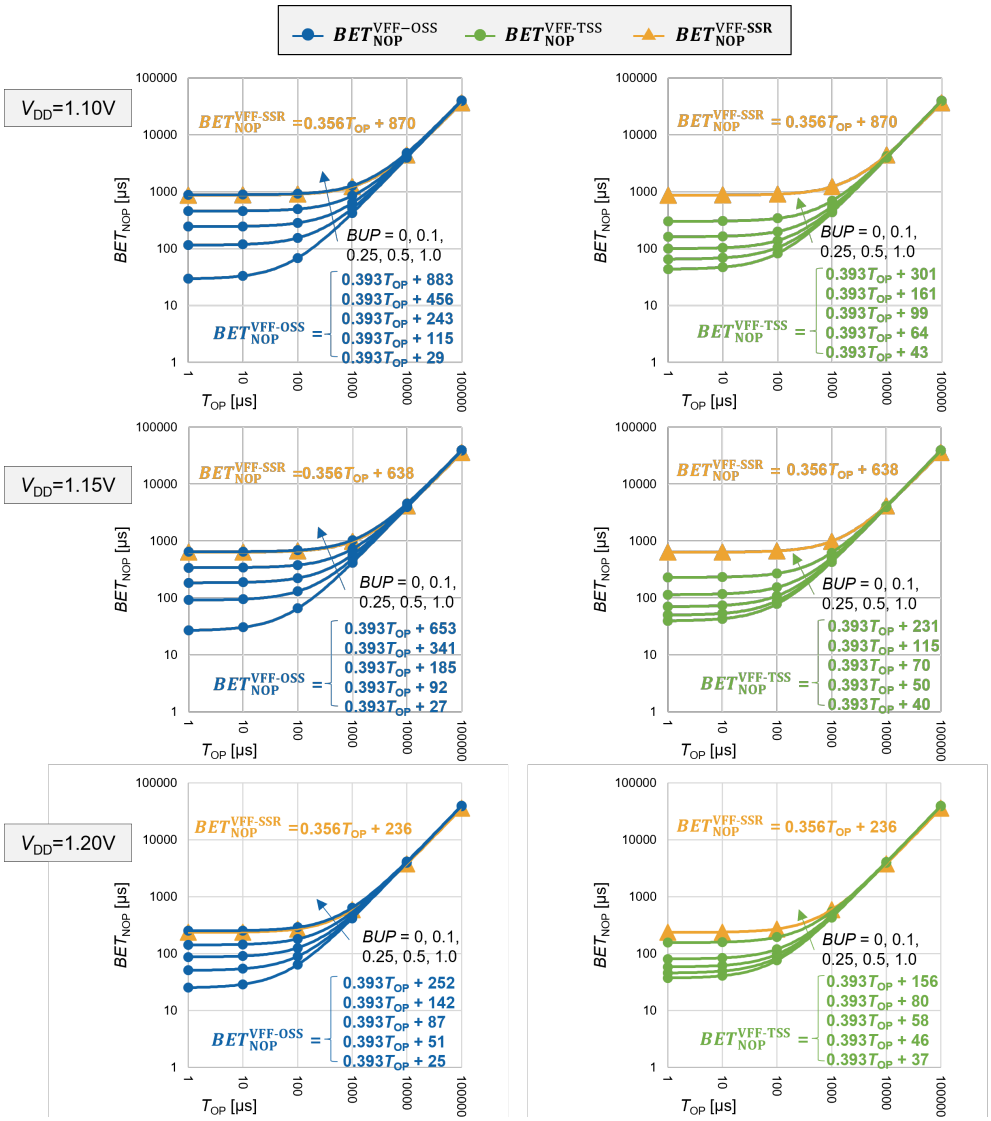


Figure 6.5. Decision making map resulted from breakeven analysis comparing VFF, SSR-NVFF and VR-NVFF using $BET_{\mathrm{NOP}}^{\mathrm{VFF\text{-}SSR}}$, $BET_{\mathrm{NOP}}^{\mathrm{VFF\text{-}OSS}}$ and $BET_{\mathrm{NOP}}^{\mathrm{VFF\text{-}TSS}}$ in various BUP and V_{DD}.

$BET_{\mathrm{NOP}}^{\mathrm{VFF\text{-}SSR}}$, $BET_{\mathrm{NOP}}^{\mathrm{VFF\text{-}OSS}}$, and $BET_{\mathrm{NOP}}^{\mathrm{VFF\text{-}TSS}}$ calculated using the pro-

posed energy model are drawn to construct the decision making maps, but on log-log graphs as illustrated in Fig. 6.5.

Each BET_{NOP} here is based on VFF, and the lower it is, the wider the range of energy reduction by NVPG. $BET_{\mathrm{NOP}}^{\text{VFF-OSS}}$ and $BET_{\mathrm{NOP}}^{\text{VFF-TSS}}$ vary with BUP, and the differences are greater for lower V_{DD}.

On a log-log graph, the slopes appear moderate when T_{OP} is relatively small, but become steeper as T_{OP} increases. Recall that BET_{NOP} consists of two components: the overhead due to leakage and the NVPG control. These changes in the steepness of BET_{NOP} in log-log graphs indicate changes in the impact of each component on BET_{NOP} with increasing T_{OP}. At $T_{\mathrm{OP}} = BET_{\mathrm{NVPG}}/\eta_{\mathrm{leak}}$, leakage energy and NVPG control overhead even ($\Leftrightarrow BET_{\mathrm{leak}} = BET_{\mathrm{NVPG}}$). When $T_{\mathrm{OP}} < BET_{\mathrm{NVPG}}/\eta_{\mathrm{leak}}$, the NVPG control energy is dominant; otherwise, the leakage energy associated with T_{OP} increase is dominant. Hence, the benefit from DAS and the TSS functionality can be expected only in applications with relatively short T_{OP}.

The intercepts of each line BET_{NOP} are BET_{NVPG} as described in Equation (6.3). The intercepts of $BET_{\mathrm{NOP}}^{\text{VFF-SSR}}$, $BET_{\mathrm{NOP}}^{\text{VFF-OSS}}$, and $BET_{\mathrm{NOP}}^{\text{VFF-TSS}}$ for various BUP and V_{DD} are shown in Fig. 6.6. For VR-NVFF, since the effect of store energy reductions depends on BUP, the intercepts of $BET_{\mathrm{NOP}}^{\text{VFF-OSS}}$ and $BET_{\mathrm{NOP}}^{\text{VFF-TSS}}$ are smaller for lower BUP. The OSS control achieves the reduction proportional to BUP due to DAS functionality but worsens compared to SSR-NVFF when BUP is high due to the overhead of verify operation. On the other hand, the TSS control remains highly effective in energy reduction even with high BUP. When $V_{\mathrm{DD}} = 1.10$ V and 1.20 V, $BET_{\mathrm{NVPG}}^{\text{VFF-TSS}}$ is reduced by 65-95 % and 34-84 %, respectively.

In Fig. 6.6, $BEBUP$ is also depicted as the X coordinate at the intersection of $BET_{\mathrm{NOP}}^{\text{VFF-OSS}}$ and $BET_{\mathrm{NOP}}^{\text{VFF-TSS}}$ at each V_{DD}. For $V_{\mathrm{DD}} = 1.10$, 1.15, and 1.20 V, $BEBUP$ are calculated to be 2.1, 2.3 and 7. 1 %, respectively. This is consistent with the result that the magnitude relations between the store energy by the OSS and the TSS control inverts when BUP is greater than 6.1 and less than 12.4 % at 1.20 V, as shown in Fig. 4.11 (b). Furthermore, the measurement results that E_{TSS} is always less than E_{OSS} at 1.10, as shown in Fig. 4.11 (a) is attributed to the fact that $BEBUP$ at 1.10 V is out of the measurement range of BUP.

Furthermore, since $\eta_{\mathrm{leak}}^{\mathrm{SSR}} < \eta_{\mathrm{leak}}^{\{\mathrm{OSS,TSS}\}}$, two lines $BET_{\mathrm{NOP}}^{\text{VFF-SSR}}$ and $BET_{\mathrm{NOP}}^{\text{VFF-}\{\mathrm{OSS,TSS}\}}$ have an intersection at certain T_{OP}. The X-coordinate of the intersection point

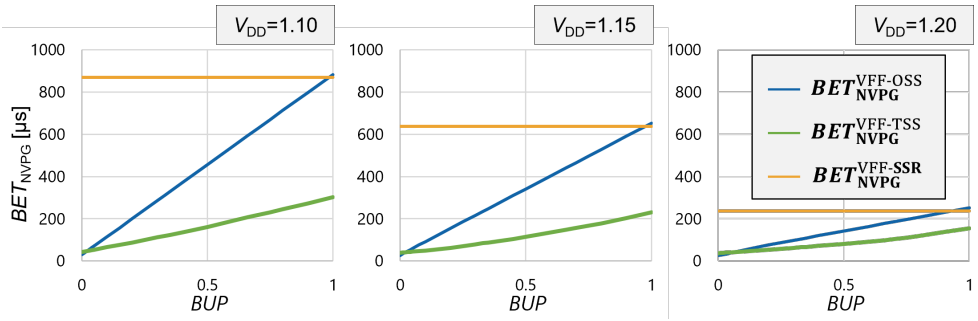


Figure 6.6. $BET_{\mathrm{NVPG}}^{\mathrm{VFF\text{-}\{SSR,OSS,TSS\}}}$: NVPG control overhead component in BET_{NOP} of MTJ-based NVFF versus VFF.

can be analytically calculated using Equation (6.7) and is shown in Fig. 6.7.

$$BET_{\mathrm{OP}}^{\mathrm{\{OSS,TSS\}\text{-}SSR}} = \frac{BET_{\mathrm{NVPG}}^{\mathrm{SSR}} - BET_{\mathrm{NVPG}}^{\mathrm{\{OSS,TSS\}}}}{\eta_{\mathrm{leak}}^{\mathrm{\{OSS,TSS\}}} - \eta_{\mathrm{leak}}^{\mathrm{SSR}}} \tag{6.7}$$

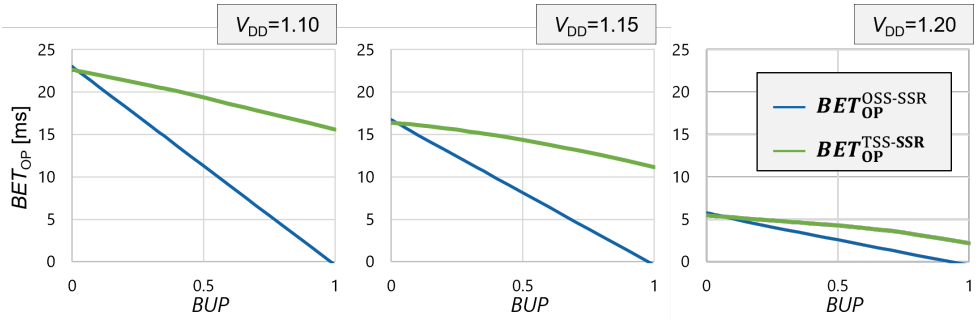


Figure 6.7. Breakeven point between SSR-NVFF and VR-NVFF: $BET_{\mathrm{OP}}^{\mathrm{\{OSS,TSS\}\text{-}SSR}}$

If the target application has T_{OP} longer than $BET_{\mathrm{OP}}^{\mathrm{\{OSS,TSS\}\text{-}SSR}}$, the system designers are recommended to choose SSR-NVFF for minimizing energy consumption. Except for the case where BUP is extremely low, BET_{OP} is longer when the TSS control is used, indicating that the TSS function of VR-NVFF is working effectively to expand the range of applications where VR-NVFF is the optimal choice.

6.2.2 MTJ-based NVFF vs. VFF vs. RFF

(a) Analysis method

RFF, a potentially good alternative, is added to the analysis in this subsection. By adding RFF into the analysis, the energy-minimizing flow are extended as shown in Fig. 6.8, and the decision making maps are made as illustrated in Fig. 6.9.

Since storing data in latches with high V_{th} transistors, store overhead of RFF is quite trivial. However, leakage current keeps consuming energy during standby states because the circuit cannot be fully power gated. Therefore, when T_{NOP} is longer than $BET_{\text{NOP}}^{\text{RFF-\{SSR,OSS,TSS\}}}$, NVFFs with zero leakage power are more energy efficient than RFF. As a results, the designer is encouraged to choose RFF when T_{NOP} is above $BET_{\text{NOP}}^{\text{VFF-RFF}}$ and below $BET_{\text{NOP}}^{\text{RFF-\{SSR,OSS,TSS\}}}$, as indicated by the purple region in Fig. 6.9.

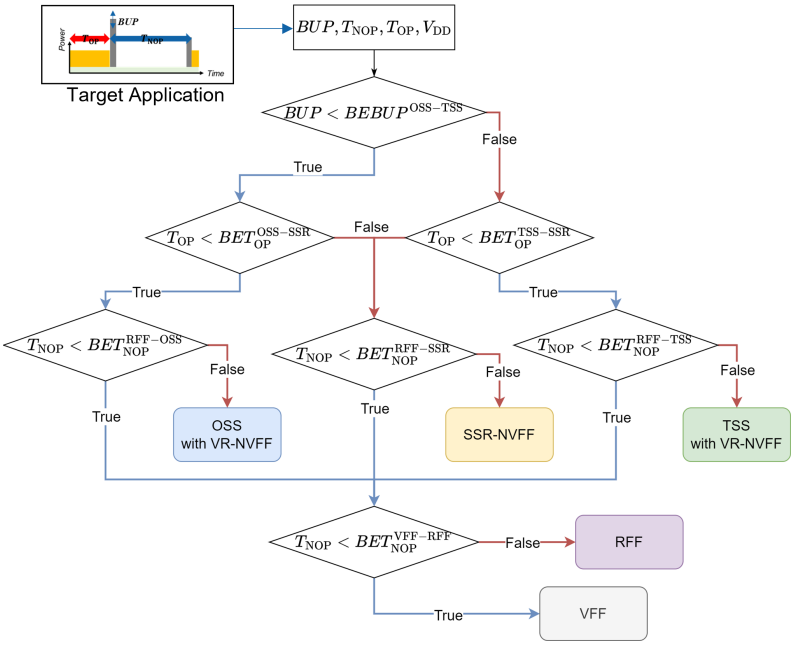


Figure 6.8. The decision flow of breakeven analysis including RFF.

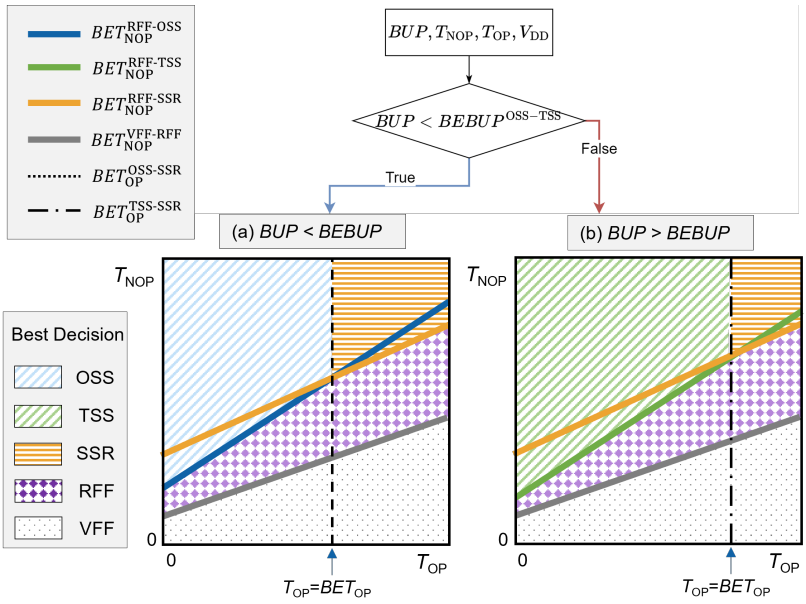


Figure 6.9. The decision making map of breakeven analysis including RFF.

(b) Analysis results

For various BUP and V_{DD}, $BET_{NOP}^{RFF-SSR}$, $BET_{NOP}^{RFF-OSS}$, $BET_{NOP}^{RFF-TSS}$, and $BET_{NOP}^{VFF-RFF}$ calculated using the proposed energy model are shown on the log-log graph shown in Fig. 6.10. $BET_{NOP}^{VFF-RFF}$ is significantly smaller compared to $BET_{NOP}^{VFF-SSR}$, $BET_{NOP}^{VFF-OSS}$, and $BET_{NOP}^{VFF-TSS}$. As the assumption in Equation (5.22) ($E_{NVC}^{RFF} \approx 0$), the intercept of $BET_{NOP}^{VFF-RFF}$ is almost zero. When T_{OP}:$T_{NOP} = 1 : 0.333$, the energy consumption of RFF is breakeven with that of VFF. Thus, it is suggested that RFF is the best solution for very fine-grained NVPG (switching in increments of 10-100 μs) below the breakeven point in MTJ-based NVFFs. However, it is true only to the extent that T_{NOP} does not exceed at most the order of 1000 μs. Otherwise, the overhead due to leakage energy during T_{NOP} becomes too significant to ignore.

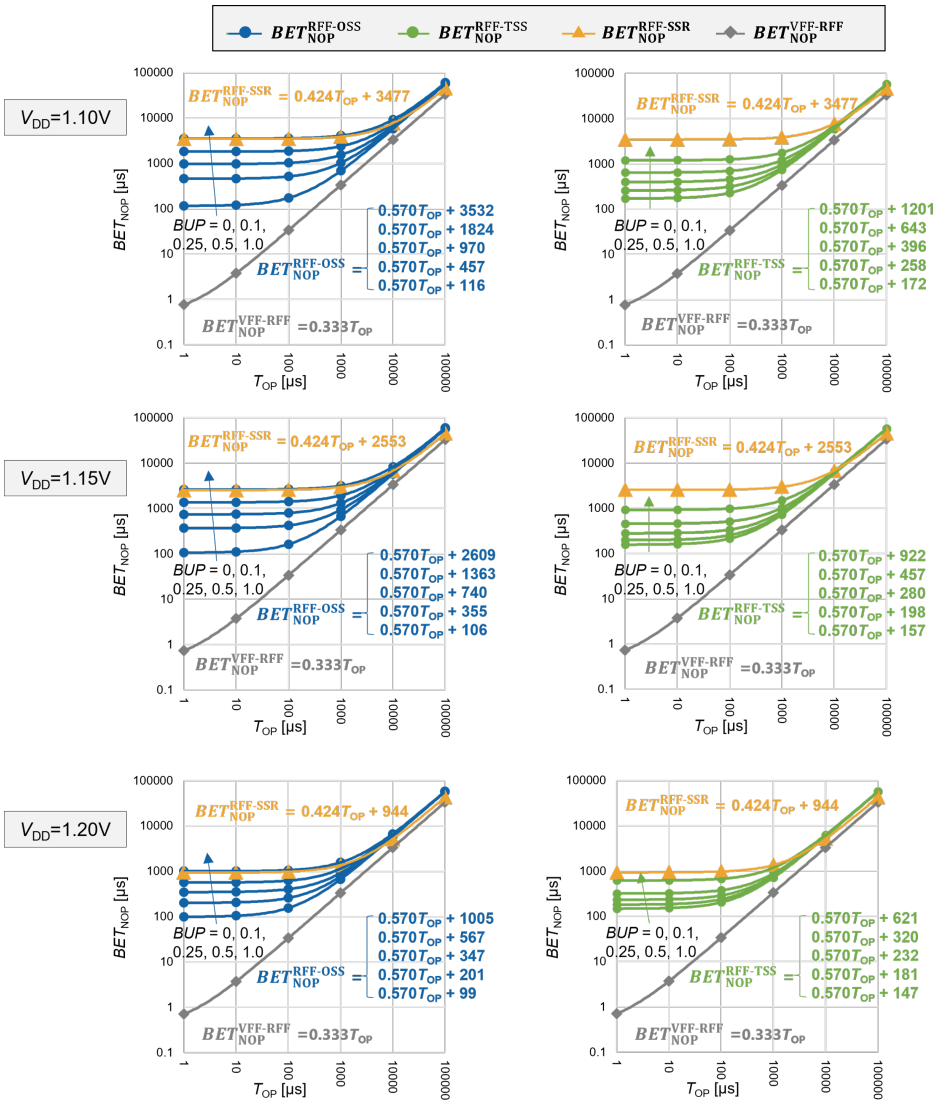


Figure 6.10. Decision making map resulted from breakeven analysis including RFF using $BET_{\mathrm{NOP}}^{\mathrm{RFF\text{-}SSR}}$, $BET_{\mathrm{NOP}}^{\mathrm{RFF\text{-}OSS}}$, $BET_{\mathrm{NOP}}^{\mathrm{RFF\text{-}TSS}}$, and $BET_{\mathrm{NOP}}^{\mathrm{VFF\text{-}RFF}}$ in various BUP and V_{DD}.

6.2.3 Discussion on dynamic voltage scaling in VR-NVFF

In the discussions so far, it has been assumed that the supply voltage V_{DD} is given based on the requirements of the application or system, and the dynamic voltage scaling (DVS) of the power supply voltage has not been considered.

However, from the chip observation results and energy model analysis, it has been suggested that MTJ-based NVFF can reduce the store energy and shorten the BET when the voltage applied to the MTJs is higher. This is believed because a sufficient high voltage for MTJ switching can achieve a high pass rate in a short store time.

While the NVCMA/MC chip does not have a DVS control mechanism, here we use an energy model to estimate the energy reduction effect of DVS. As defined in Chapter 5, NVFF energy models do not consider the peripheral circuits for control, but only for the cells, so the overhead of the control circuit is not included in the evaluation. It should be noted that this evaluation is optimistic, ignoring all temporal, control, and circuit overheads associated with the introduction and execution of DVS. In the evaluation, it is assumed that V_{DD} during normal operation is fixed at 1.10 V, and the change in BET_{NVPG} (the NVPG control overhead component of BET_{NOP}) is compared when V_{DD} increases from 1.10 V to either 1.15 V or 1.20 V for NVPG control, as shown in Fig. 6.11. The voltage during T_{OP} is always set to 1.10 V, so BET_{leak} is not affected.

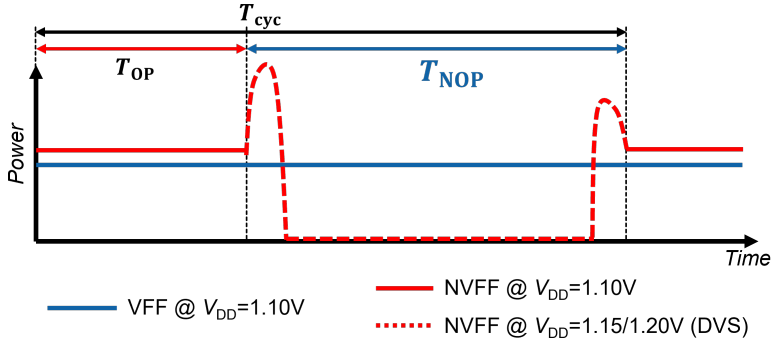


Figure 6.11. Power transition in DVS scenario during NVPG.

Calculated BET_{NVPG} is shown in Fig. 6.12 while Fig. 6.13 represents the normalized BET_{NVPG} when BUP is 0.5.

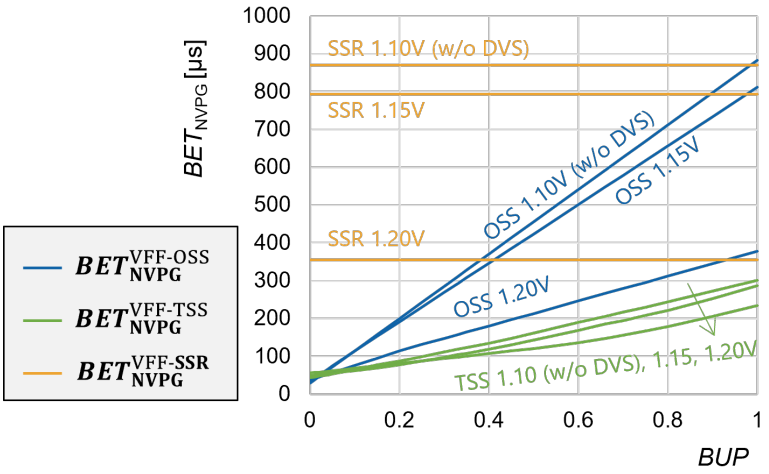


Figure 6.12. $BET_{\mathrm{NVPG}}^{\mathrm{VFF\text{-}SSR}}$, $BET_{\mathrm{NVPG}}^{\mathrm{VFF\text{-}OSS}}$: NVPG control overhead component in BET_{NOP} of MTJ-based NVFF versus VFF considering DVS for NVPG control.

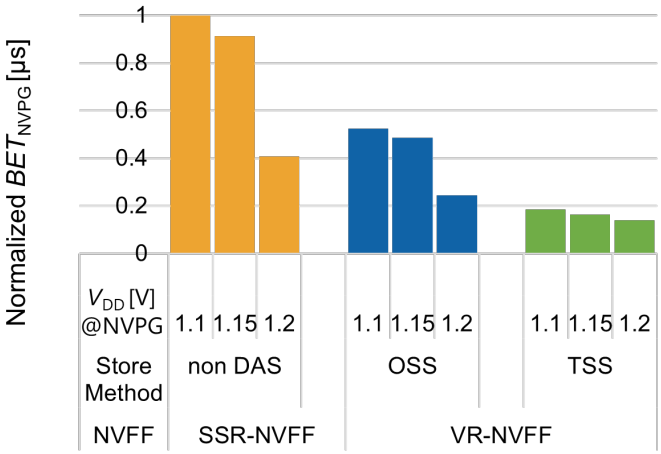


Figure 6.13. Normalised $BET_{\mathrm{NVPG}}^{\mathrm{VFF\text{-}SSR}}$, $BET_{\mathrm{NVPG}}^{\mathrm{VFF\text{-}OSS}}$, and $BET_{\mathrm{NVPG}}^{\mathrm{VFF\text{-}TSS}}$ considering DVS for NVPG control at $BUP = 0.5$.

Regardless of the store method, dynamical scaling of the voltage for NVPG control improves NVPG efficiency and shortens BET_{NOP}. When scaling from 1.10 V to 1.20 V, the effect of DVS reduces BET_{NVPG} by 59.2% for SSR-

NVFF, 27.9% for the OSS control, and 4.6% for the TSS control. In particular in SSR-NVFF and the OSS control, the reduction effect on BET_{NVPG} was greater when increasing from 1.15 V to 1.20 V than from 1.10 V to 1.15 V.

However, what should be noted here is the high energy efficiency of the TSS control. The fact that the BET_{NOP} reduction effect achieved through TSS control exceeds that achieved by DVS underscores that TSS control is an effective energy reduction option.

6.3 Summary

In this chapter, a comprehensive breakeven analysis is performed to compare various FF technologies using a proposed energy model. The analysis aims to assess the balance between the energy savings and overheads of different technologies, such as volatile FF, SSR-NVFF, VR-NVFF, and retention FF, to assist system designers in their decision-making process for energy minimization.

The chapter defines breakeven points using indicators such as T_{OP}, T_{NOP}, and BUP, which are key metrics to define intermittent operation applications. These metrics are critical to determine the most efficient FF technology for specific scenarios. The analysis presents decision-making maps to select the optimal FF technology under various conditions. For instance, the breakeven analysis shows that VR-NVFF with the TSS control significantly reduces store energy, especially at high BUP values, making it a preferred choice for applications with shorter T_{OP}s. Decision-making maps and analysis results highlight the importance of choosing the right technology based on the application's T_{OP}, T_{NOP}, and BUP parameters. Furthermore, the chapter discusses RFF as a potential alternative, especially in cases with fine-grained NVPG. However, it also notes that the advantages of RFF diminish with longer T_{NOP} due to its inability to fully power gate its circuits.

The chapter also discusses the impact of DVS on the NVPG efficiency. The results suggest that dynamically scaling the voltage for NVPG control significantly reduces the breakeven point overhead. However, the results also show that TSS control can achieve a sufficiently high energy reduction effect without using DVS, which involves the complexity of multi-V_{DD} design, supporting the effectiveness of the TSS control.

In summary, Chapter 6 provides a detailed and practical workflow for system designers to evaluate and select the most energy-efficient FF technology for their specific applications. Employing a comprehensive breakeven analysis, the chapter provides valuable insights on the relative advantages of different

FF technologies in terms of energy minimization, particularly in intermittent operation applications.

7

Conclusion and Future Work

7.1 Conclusion

The significance of low-power LSIs for devices in edge computing is increasingly recognized. For the design of energy-efficient chips, it is crucial to estimate the power consumption, especially for target applications, beforehand and select the optimal technology accordingly. Over the past decades, while semiconductor miniaturization has enhanced computational performance, the leak current in transistors has also increased, becoming a non-negligible part of power consumption. This research aims to assist system designers in considering the application of NVPG, a technique to reduce leak power during idle periods without data loss, for intermittent operation applications in edge computing.

Firstly, this thesis analyzes the variability in switching delay times of STT-MTJs and demonstrates the effectiveness of the TSS control energy reduction technique in VR-NVFF, based on actual measurements of a 40 nm MTJ/C-MOS hybrid chip accelerator. These results underscore the necessity for an energy model that can estimate energy under various conditions to maximize energy efficiency.

Next, based on the actual measurements, we propose an energy model for VR-NVFF targeting intermittent operation applications by modeling the MTJ variability characteristics as a pass rate model assuming to follow the cumulative distribution function of a Gaussian distribution. This model has been validated against actual measurements, proving its adequacy in determining

the optimal store method and store timing for the TSS control under arbitrary conditions.

Furthermore, we have defined energy models for ordinary volatile FFs, retention FFs with high V_{th}, and SSR-NVFFs as well. Leveraging these models, we introduce a breakeven analysis methodology to facilitate the quantitative assessment of different FF technologies. This method provides quantitative criteria for evaluating the balance between the energy reduction effects and overheads of various alternatives, thereby guiding the selection of the most energy-efficient FF technology for the target application. The contributions of this thesis are all encapsulated in this comprehensive decision-making workflow.

7.2 Future work

This work proposes an energy model based on data obtained from experiments using actual chips, and assumes a scenario in which design exploration is conducted under the same conditions. However, the cost of MTJ switching is greatly influenced by its characteristics and is in a trade-off relationship with retention time [73]. If the target application assumes fine-grained PG, which is the specialty of RFF, optimizing the MTJ characteristics so that the switching cost is small while the retention time is short can improve the BET_{NOP} of MTJ-based NVFF. Therefore, investigating the change in the energy model of NVFF due to the characteristics of MTJ is a meaningful future work. The proposed model is able to reflect such a diversity of MTJ characteristics in the pass rate model, and the power required for MTJ switching is also parameterized in this model. Thus, the model can be extended to model variety NVFFs with different MTJ characteristics.

Furthermore, the model proposed in this work assumes operation in an environment with relatively moderate temperature change (from -10 to 100°C). The extension of the model to cover more extreme temperature environments where edge devices are possibly installed, especially low temperatures where MTJ is less likely to switch (but the switching current is more likely to flow), is meaningful to expand the applicability of the model.

In addition, the incorporation of our contributions in this thesis into circuit simulators and VLSI design CAD tools is also a meaningful future work. For instance, by integrating the proposed energy model into a simulator such as the cycle-accurate simulator CubeSim [83] realizes the combination of the solid energy estimation with a measurement-based model and the flexibility of the simulator to freely modify the memory designs. This will facilitate the design exploration of energy-efficient NVFF-embedded architectures and the further

development of edge computing applications.

Lastly, NVCMA/MC is equipped with ECC with high error correction strength in trades for yield improvement, but it is also meaningful to consider introducing ECC for the purpose of energy reduction, for example, by omitting the second store in TSS control to achieve energy reduction.

Acknowledgement

本論文の完成にあたり、多大なるご支援と温かいご指導を賜りました天野英晴教授に心から感謝申し上げます。研究に行き詰った時、常に前向きなお言葉に助けられました。定年退職を迎えられる最後の年に、ふんがさんのもとで研究できたことを心から光栄に思います。

東京大学助教の小島拓也先生には、先輩として研究の初期段階から厚いご指導と的確なアドバイスをいただき感謝の気持ちでいっぱいです。博士課程の学生として、また研究者としての姿勢を学ばせていただきました。

共同研究を進める中で、多くの知見と探求の機会を提供してくださった芝浦工業大学の宇佐美公良教授及び宇佐美研究室の小野義基さん、宮内陽里さん、横山大輝さんにも深く感謝申し上げます。 皆様との議論は、私の研究にとって大きなモチベーションでした。

共同研究先であるソニーセミコンダクタソリューションズ株式会社の別所和宏さん、平賀啓三さん、鈴木健太さんにも厚く御礼を申し上げます。実務に基づく具体的な助言は、学術的な研究にとどまらず、実社会での応用についての考え方を深め、研究の意義を再認識する機会となりました。

卒業生を含めた天野研究室のメンバーの皆様にも心から感謝申し上げます。皆様が積み上げられた研究成果と、受け継がれた設備・環境があったからこそ、不自由なく研究を進めることができました。 特に、現東京農工大学助教の丹羽直也さん、最後まで共に闘ってくれた飯塚健介さんの献身的な研究室運営や後輩指導がなければ、天野研究室での研究は成り立たなかったと思います。心から感謝申し上げます。

進学に際して背中を押していただいた当時の誘実隊長 田中 1 佐及び航空幕僚監部 木下 1 佐、各種調整に尽力いただいた江刺 3 佐並びに研究の進捗を見守ってくださった小原 3 佐、丸山 3 佐をはじめとする航空自衛隊の皆様にも感謝致します。天野研究室をご紹介いただいた防衛大学校の黒川恭一教授及び岩井啓輔准教授にも多大なる感謝を申し上げます。

最後に、私の研究活動を支えてくださった家族、友人及び友人の猫に心から感謝します。 ありがとうございました。

Aika Kamei
February 2024

Bibliography

[1] S. Shigematsu, S. Mutoh, Y. Matsuya, Y. Tanabe, and J. Yamada, "A 1-v high-speed mtcmos circuit scheme for power-down application circuits," *IEEE Journal of Solid-State Circuits*, vol. 32, no. 6, pp. 861–869, 1997.

[2] H. Mahmoodi-Meimand and K. Roy, "Data-retention flip-flops for power-down applications," in *2004 IEEE International Symposium on Circuits and Systems (IEEE Cat. No. 04CH37512)*, vol. 2. IEEE, 2004, pp. II–677.

[3] H. Jiao and Z. Zhang, "A compact low-power data retention flip-flop with easy-sleep mode," in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2020, pp. 1–5.

[4] S. Sedighiani, K. Singh, R. Jordans, P. Harpe, and J. P. de Gyvez, "A 380 fw leakage data retention flip-flop for short sleep periods," *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2023.

[5] N. S. Kim, T. Austin, D. Baauw, T. Mudge, K. Flautner, J. S. Hu, M. J. Irwin, M. Kandemir, and V. Narayanan, "Leakage current: Moore's law meets static power," *computer*, vol. 36, no. 12, pp. 68–75, 2003.

[6] J. Li, C. Augustine, S. Salahuddin, and K. Roy, "Modeling of failure probability and statistical design of spin-torque transfer magnetic random access memory (stt mram) array for yield enhancement," in *Proceedings of the 45th annual Design Automation Conference*, 2008, pp. 278–283.

[7] Y. Zhang, X. Wang, and Y. Chen, "Stt-ram cell design optimization for persistent and non-persistent error rate reduction: A statistical design view," in *2011 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*. IEEE, 2011, pp. 471–477.

[8] A. Nigam, C. W. Smullen, V. Mohan, E. Chen, S. Gurumurthi, and M. R. Stan, "Delivering on the promise of universal memory for spin-transfer

torque ram (stt-ram)," in *IEEE/ACM International Symposium on Low Power Electronics and Design*. IEEE, 2011, pp. 121–126.

[9] P. Wang, W. Zhang, R. Joshi, R. Kanj, and Y. Chen, "A thermal and process variation aware mtj switching model and its applications in soft error analysis," in *Proceedings of the International Conference on Computer-Aided Design*, 2012, pp. 720–727.

[10] C. Jiang, T. Fan, H. Gao, W. Shi, L. Liu, C. Cérin, and J. Wan, "Energy aware edge computing: A survey," *Computer Communications*, vol. 151, pp. 556–580, 2020.

[11] H. Jayakumar, A. Raha, Y. Kim, S. Sutar, W. S. Lee, and V. Raghunathan, "Energy-efficient system design for iot devices," in *2016 21st Asia and South Pacific design automation conference (ASP-DAC)*. IEEE, 2016, pp. 298–301.

[12] W. Yang and H. Thapliyal, "Low-power and energy-efficient full adders with approximate adiabatic logic for edge computing," in *2020 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*. IEEE, 2020, pp. 312–315.

[13] C. Chen, S. Yang, W. Qian, M. Imani, X. Yin, and C. Zhuo, "Optimally approximated and unbiased floating-point multiplier with runtime configurability," in *Proceedings of the 39th international conference on computer-aided design*, 2020, pp. 1–9.

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

[15] arm Developer, "Arm musca-s1 test chip and board technical overview." accessed on Jan 24th, 2024. [Online]. Available: https://developer.arm.com/documentation/101756/0000/introduction/the-musca-s1-development-board-at-a-glance

[16] Y. Shuto, S. Yamamoto, and S. Sugahara, "Comparative study of power-gating architectures for nonvolatile finfet-sram using spintronics-based retention technology," in *2015 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2015, pp. 866–871.

[17] H. J. Veendrick, "Short-circuit dissipation of static cmos circuitry and its impact on the design of buffer circuits," *IEEE Journal of Solid-State Circuits*, vol. 19, no. 4, pp. 468–473, 1984.

[18] A. Hirata, H. Onodera, and K. Tamaru, "Estimation of short-circuit power dissipation and its influence on propagation delay for static cmos gates," in *1996 IEEE International Symposium on Circuits and Systems. Circuits and Systems Connecting the World. ISCAS 96*, vol. 4. IEEE, 1996, pp. 751–754.

[19] K. Nose and T. Sakurai, "Analysis and future trend of short-circuit power," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 19, no. 9, pp. 1023–1030, 2000.

[20] N. H. Weste and D. Harris, *CMOS VLSI design: a circuits and systems perspective*. Pearson Education India, 2011.

[21] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer cmos circuits," *Proceedings of the IEEE*, vol. 91, no. 2, pp. 305–327, 2003.

[22] M. Horowitz, E. Alon, D. Patil, S. Naffziger, R. Kumar, and K. Bernstein, "Scaling, power, and the future of cmos," in *IEEE InternationalElectron Devices Meeting, 2005. IEDM Technical Digest.* IEEE, 2005, pp. 7–pp.

[23] J.-O. Seo, M. Seok, and S. Cho, "Archon: A 332.7 tops/w 5b variation-tolerant analog cnn processor featuring analog neuronal computation unit and analog memory," in *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 65. IEEE, 2022, pp. 258–260.

[24] R. M. Rao, C. Gonzalez, E. Fluhr, A. Mathews, A. Bianchi, D. Dreps, D. Wolpert, E. Lai, G. Strevig, G. Wiedemeier *et al.*, "Power10™: A 16-core smt8 server processor with 2tb/s off-chip bandwidth in 7nm technology," in *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 65. IEEE, 2022, pp. 48–50.

[25] D. Kim, N. M. Rahman, and S. Mukhopadhyay, "29.1 a 32.5 mw mixed-signal processing-in-memory-based k-sat solver in 65nm cmos with 74.0% solvability for 3d-variable 126-clause 3-sat problems," in *2023 IEEE International Solid-State Circuits Conference (ISSCC)*. IEEE, 2023, pp. 28–30.

[26] I. T. I. R. F. DEVICES and SYSTEMS, "International roadmap for devices and systems 2022 edition." 2022, accessed on Jan 24th, 2024. [Online]. Available: https://www.yolegroup.com/product/report/emerging-non-volatile-memory-2023/

[27] B. Pourshirazi, M. V. Beigi, Z. Zhu, and G. Memik, "Writeback-aware llc management for pcm-based main memory systems," *ACM Transactions on Design Automation of Electronic Systems (TODAES)*, vol. 24, no. 2, pp. 1–19, 2019.

[28] Intel, "Intel optane technology," accessed on Jan 24th, 2024. [Online]. Available: https://www.intel.com/content/www/us/en/architecture-and-technology/optane-technology/optane-for-data-centers.html

[29] Y. Intelligence, "Emerging non-volatile memory 2023." 2023, accessed on Jan 24th, 2024. [Online]. Available: https://irds.ieee.org/images/files/pdf/2022/2022IRDS_BC.pdf

[30] S. Yamamoto and S. Sugahara, "Nonvolatile delay flip-flop based on spin-transistor architecture and its power-gating applications," *Japanese Journal of Applied Physics*, vol. 49, no. 9R, p. 090204, 2010.

[31] S. Yamamoto, Y. Shuto, and S. Sugahara, "Nonvolatile delay flip-flop using spin-transistor architecture with spin transfer torque mtjs for power-gating systems," *Electronics letters*, vol. 47, no. 18, pp. 1027–1029, 2011.

[32] Y. Shuto, R. Nakane, W. Wang, H. Sukegawa, S. Yamamoto, M. Tanaka, K. Inomata, and S. Sugahara, "A new spin-functional metal–oxide–semiconductor field-effect transistor based on magnetic tunnel junction technology: Pseudo-spin-mosfet," *Applied physics express*, vol. 3, no. 1, p. 013003, 2010.

[33] S. Sugahara and M. Tanaka, "A spin metal–oxide–semiconductor field-effect transistor using half-metallic-ferromagnet contacts for the source and drain," *Applied Physics Letters*, vol. 84, no. 13, pp. 2307–2309, 2004.

[34] S. Lee, H. Koike, M. Goto, S. Miwa, Y. Suzuki, N. Yamashita, R. Ohshima, E. Shigematsu, Y. Ando, and M. Shiraishi, "Synthetic rashba spin–orbit system using a silicon metal-oxide semiconductor," *Nature Materials*, vol. 20, no. 9, pp. 1228–1232, 2021.

[35] T. Endo, S. Tsuruoka, Y. Tadano, S. Kaneta-Takada, Y. Seki, M. Kobayashi, L. D. Anh, M. Seki, H. Tabata, M. Tanaka *et al.*, "Giant spin-valve effect in planar spin devices using an artificially implemented nanolength mott-insulator region," *Advanced Materials*, p. 2300110, 2023.

[36] M. Kudo and K. Usami, "Nonvolatile power gating with mtj based non-volatile flip-flops for a microprocessor," in *2017 IEEE 6th Non-Volatile Memory Systems and Applications Symposium (NVMSA)*. IEEE, 2017, pp. 1–6.

[37] T. Ikezoe, H. Amano, J. Akaike, K. Usami, M. Kudo, K. Hiraga, Y. Shuto, and K. Yagami, "A coarse grained-reconfigurable accelerator with energy efficient mtj-based non-volatile flip-flops," in *2018 International Conference on ReConFigurable Computing and FPGAs, ReConFig 2018, Cancun, Mexico, December 3-5, 2018*, 2018, pp. 1–6.

[38] K. Usami, J. Akaike, S. Akiba, M. Kudo, H. Amano, T. Ikezoe, K. Hiraga, Y. Shuto, and K. Yagami, "Energy efficient write verify and retry scheme for mtj based flip-flop and application," in *2018 IEEE 7th Non-Volatile Memory Systems and Applications Symposium (NVMSA)*. IEEE, 2018, pp. 91–98.

[39] M. Kudo and K. Usami, "Mtj based non-volatile flip flop to prevent useless store operation," in *ITC-CSCC: International Technical Conference on Circuits Systems, Computers and Communications*, 2015, pp. 515–518.

[40] Y. Zhang, W. Zhao, G. Prenat, T. Devolder, J. Klein, C. Chappert, B. Dieny, and D. Ravelosona, "Electrical modeling of stochastic spin transfer torque writing in magnetic tunnel junctions for memory and logic applications," *IEEE Transactions on Magnetics*, vol. 49, no. 7, pp. 4375–4378, July 2013.

[41] A. F. Vincent, N. Locatelli, J. Klein, W. S. Zhao, S. Galdin-Retailleau, and D. Querlioz, "Analytical Macrospin Modeling of the Stochastic Switching Time of Spin-Transfer Torque Devices," *IEEE Transactions on Electron Devices*, vol. 62, no. 1, pp. 164–170, Jan 2015.

[42] R. De Rose, M. Lanuzza, F. Crupi, G. Siracusano, R. Tomasello, G. Finocchio, M. Carpentieri, and M. Alioto, "A variation-aware timing modeling approach for write operation in hybrid cmos/stt-mtj circuits," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 65, no. 3, pp. 1086–095, March 2018.

[43] K. Usami, M. Igarashi, F. Minami, T. Ishikawa, M. Kanzawa, M. Ichida, and K. Nogami, "Automated low-power technique exploiting multiple supply voltages applied to a media processor," *IEEE Journal of Solid-State Circuits*, vol. 33, no. 3, pp. 463–472, 1998.

[44] S. H. Kulkarni and D. Sylvester, "Power distribution techniques for dual vdd circuits," in *Proceedings of the 2006 Asia and South Pacific Design Automation Conference*, 2006, pp. 838–843.

[45] Y. Shuto, S. Sugahara *et al.*, "Nonvolatile flip-flop based on pseudo-spin-transistor architecture and its nonvolatile power-gating applications for

low-power cmos logic," *The European Physical Journal Applied Physics*, vol. 63, no. 1, p. 14403, 2013.

[46] D. Chabi, W. Zhao, E. Deng, Y. Zhang, N. B. Romdhane, J.-O. Klein, and C. Chappert, "Ultra low power magnetic flip-flop based on checkpointing/power gating and self-enable mechanisms," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 61, no. 6, pp. 1755–1765, 2014.

[47] D. Suzuki and T. Hanyu, "Magnetic-tunnel-junction based low-energy nonvolatile flip-flop using an area-efficient self-terminated write driver," *Journal of Applied Physics*, vol. 117, no. 17, 2015.

[48] M. Kazemi, E. Ipek, and E. G. Friedman, "Energy-efficient nonvolatile flip-flop with subnanosecond data backup time for fine-grain power gating," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 62, no. 12, pp. 1154–1158, 2015.

[49] A. S. Iyengar, S. Ghosh, and J.-W. Jang, "Mtj-based state retentive flip-flop with enhanced-scan capability to sustain sudden power failure," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 62, no. 8, pp. 2062–2068, 2015.

[50] R. Bishnoi, F. Oboril, and M. B. Tahoori, "Fault tolerant non-volatile spintronic flip-flop," in *2016 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2016, pp. 261–264.

[51] N. Onizawa, A. Mochizuki, A. Tamakoshi, and T. Hanyu, "Sudden power-outage resilient in-processor checkpointing for energy-harvesting nonvolatile processors," *IEEE Transactions on Emerging Topics in Computing*, vol. 5, no. 2, pp. 151–163, 2016.

[52] M. Lanuzza, R. De Rose, F. Crupi, and M. Alioto, "An energy aware variation-tolerant writing termination control for stt-based non volatile flip-flops," in *2019 26th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*. IEEE, 2019, pp. 158–161.

[53] K. Monga, N. Chaturvedi, and S. Gurunarayanan, "Energy-efficient data retention in d flip-flops using stt-mtj," *Circuit World*, vol. 46, no. 4, pp. 229–241, 2020.

[54] Y. ABE, K. KOBAYASHI, J. SHIOMI, and H. OCHI, "Nonvolatile storage cells using ficc for iot processors with intermittent operations," *IEICE Transactions on Electronics*, p. 2022CTP0001, 2023.

[55] A. Lee, C.-P. Lo, C.-C. Lin, W.-H. Chen, K.-H. Hsu, Z. Wang, F. Su, Z. Yuan, Q. Wei, Y.-C. King *et al.*, "A reram-based nonvolatile flip-flop with self-write-termination scheme for frequent-off fast-wake-up nonvolatile processors," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 8, pp. 2194–2207, 2017.

[56] N. Onizawa and T. Hanyu, "Redundant stt-mtj-based nonvolatile flip-flops for low write-error-rate operations," in *2016 14th IEEE International New Circuits and Systems Conference (NEWCAS)*. IEEE, 2016, pp. 1–4.

[57] D. Suzuki and T. Hanyu, "Design of a low-power nonvolatile flip-flop using three-terminal magnetic-tunnel-junction-based self-terminated mechanism," *Japanese journal of applied physics*, vol. 56, no. 4S, p. 04CN06, 2017.

[58] "Design of a highly reliable nonvolatile flip-flop incorporating a common-mode write error detection capability."

[59] S. NAKABEPPU and N. YAMASAKI, "Non-stop microprocessor for fault-tolerant real-time systems," *IEICE Transactions on Electronics*, p. 2022CDP0005, 2023.

[60] K. Usami, S. Akiba, H. Amano, T. Ikezoe, K. Hiraga, K. Suzuki, and Y. Kanda, "Non-volatile coarse grained reconfigurable array enabling two-step store control for energy minimization," in *2020 IEEE Symposium in Low-Power and High-Speed Chips (COOL CHIPS)*. IEEE, 2020, pp. 1–3.

[61] A. Kamei, T. Kojima, H. Amano, D. Yokoyama, H. Miyauchi, K. Usami, K. Hiraga, K. Suzuki, and K. Bessho, "Energy saving in a multi-context coarse grained reconfigurable array with non-volatile flip-flops," in *2021 IEEE 14th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC)*. IEEE, 2021, pp. 273–280.

[62] N. Ozaki, Y. Yasuda, M. Izawa, Y. Saito, D. Ikebuchi, H. Amano, H. Nakamura, K. Usami, M. Namiki, and M. Kondo, "Cool mega-arrays: Ultralow-power reconfigurable accelerator chips," *IEEE Micro*, vol. 31, no. 6, pp. 6–18, Nov 2011.

[63] H. Schmit, D. Whelihan, A. Tsai, M. Moe, B. Levine, and R. R. Taylor, "PipeRench: A virtualized programmable datapath in 0.18 micron technology," in *Custom Integrated Circuits Conference, 2002. Proceedings of the IEEE 2002*. IEEE, 2002, pp. 63–66.

[64] G. Ansaloni, P. Bonzini, and L. Pozzi, "EGRA: A coarse grained reconfigurable architectural template," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 19, no. 6, pp. 1062–1074, 2011.

[65] V. Govindaraju, C.-H. Ho, T. Nowatzki, J. Chhugani, N. Satish, K. Sankaralingam, and C. Kim, "Dyser: Unifying functionality and parallelism specialization for energy-efficient computing," *IEEE Micro*, vol. 32, no. 5, pp. 38–51, 2012.

[66] G. Gobieski, A. O. Atli, K. Mai, B. Lucia, and N. Beckmann, "Snafu: an ultra-low-power, energy-minimal cgra-generation framework and architecture," in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2021, pp. 1027–1040.

[67] B. Levine, "Kilocore: Scalable, High Performance and Power Efficient Coarse Grained Reconfigurable Fabrics," in Proc. of International Symposium on Advanced Reconfigurable Systems, 2005, pp. 129–158.

[68] J. M. Arnold, "S5: The Architecture and Development Flow of a Software Configurable Processor," in *Proc. of the 4th IEEE Int'l Conf. on Field Programmable Technology (ICFPT2005)*, December 2005, pp. 120–128.

[69] T. Nowatzki, V. Gangadhar, N. Ardalani, and K. Sankaralingam, "Stream-dataflow acceleration," in *2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2017, pp. 416–429.

[70] T. IKEZOE, T. KOJIMA, and H. AMANO, "Reconvering faulty nonvolatile flip flops for coarse-grained reconfigurable architectures," *IEICE Transactions on Electronics*, 2020.

[71] A. Ohwada, T. Kojima, and H. Amano, "Mentai: A fully automated cgra application development environment that supports hardware/software co-design," in *Proceedings of SASIMI 2021*, March 2021.

[72] T. Kojima, N. A. V. Doan, and H. Amano, "Genmap: A genetic algorithmic approach for optimizing spatial mapping of coarse-grained reconfigurable architectures," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 28, no. 11, pp. 2383–2396, 2020.

[73] M. Oka, Y. Namba, Y. Sato, H. Uchida, T. Doi, T. Tatsuno, M. Nakazawa, A. Tamura, R. Haga, M. Kuroda *et al.*, "3d stacked cis compatible 40nm embedded stt-mram for buffer memory," in *2021 Symposium on VLSI Technology*. IEEE, 2021, pp. 1–2.

[74] H. LTD., "Humandata xcm-208a (japanese)," accessed on Jan 24th, 2024. [Online]. Available: https://www.hdl.co.jp/XCM-208/

[75] M. Inc., "Ldo regulator mic69502," 2006, accessed on Jan 24th, 2024. [Online]. Available: https://www.mouser.jp/datasheet/2/268/mic69502-1082391.pdf

[76] X. Bi, H. Li, and X. Wang, "Stt-ram cell design considering cmos and mtj temperature dependence," *IEEE Transactions on Magnetics*, vol. 48, no. 11, pp. 3821–3824, 2012.

[77] B. Wu, Y. Cheng, J. Yang, A. Todri-Sanial, and W. Zhao, "Temperature impact analysis and access reliability enhancement for 1t1mtj stt-ram," *IEEE Transactions on Reliability*, vol. 65, no. 4, pp. 1755–1768, 2016.

[78] Y. Wang, Y. Zhang, E. Deng, J.-O. Klein, L. A. Naviner, and W. Zhao, "Compact model of magnetic tunnel junction with stochastic spin transfer torque switching for reliability analyses," *Microelectronics Reliability*, vol. 54, no. 9-10, pp. 1774–1778, 2014.

[79] Y. Zhang, B. Yan, W. Kang, Y. Cheng, J.-O. Klein, Y. Zhang, Y. Chen, and W. Zhao, "Compact model of subvolume mtj and its design application at nanoscale technology nodes," *IEEE Transactions on Electron Devices*, vol. 62, no. 6, pp. 2048–2055, 2015.

[80] D. Worledge, G. Hu, D. W. Abraham, J. Sun, P. Trouilloud, J. Nowak, S. Brown, M. Gaidis, E. O'sullivan, and R. Robertazzi, "Spin torque switching of perpendicular ta— cofeb— mgo-based magnetic tunnel junctions," *Applied physics letters*, vol. 98, no. 2, p. 022501, 2011.

[81] E. Eken, L. Song, I. Bayram, C. Xu, W. Wen, Y. Xie, and Y. Chen, "Nvsim-vx s: an improved nvsim for variation aware stt-ram simulation," in *2016 53nd ACM/EDAC/IEEE Design Automation Conference (DAC)*. IEEE, 2016, pp. 1–6.

[82] N. S. University, "Freepdk45," accessed on Jan 24th, 2024. [Online]. Available: https://eda.ncsu.edu/freepdk/freepdk45/

[83] T. Kojima, T. Ikezoe, and H. Amano, "Cubesim: A cycle accurate simulator for multicore system with 3d sip (in japanese)," *D - Abstracts of IEICE TRANSACTIONS on Information and Systems*, vol. 104, no. 4, pp. 228–241, 2021.

Publications

Related Papers

Journal Papers

[1] Aika Kamei, Hideharu Amano, Takuya Kojima, Daiki Yokoyama, Kimiyoshi Usami, Keizo Hiraga, Kenta Suzuki, and Kazuhiro Bessho, A variation-aware MTJ store energy estimation model for edge devices with verify-and-retryable nonvolatile flip-flops, *IEEE Transactions on Very Large Scale Integration Systems*, vol. 31, no. 4, pp.532–542, April 2023.

International Conference Papers

[2] Aika Kamei, Takuya Kojima, Hideharu Amano, Daiki Yokoyama, Hisato Miyauhi, Kimiyoshi Usami, Keizo Hiraga, Kenta Suzuki, and Kazuhiro Bessho, Energy saving in a multi-context coarse grained reconfigurable array with non-volatile flip-flops, In *Proc. of 2021 IEEE 14th International Symposium on Embedded Multicore/Many-core Systems-on-Chip (MCSoC)*, pp.273–280, Singapore, December 2021.

Other Papers

Journal Papers

[3] Kimiyoshi Usami, Daiki Yokoyama, Aika Kamei, Hideharu Amano, Kenta Suzuki, Keizo Hiraga, and Kazuhiro Bessho, Optimized Two-Step Store Control for MTJ-Based Nonvolatile Flip-Flops to Minimize Store Energy Under Process and Temperature Variations, *IEEE Transactions on Very Large Scale Integration Systems*, vol. 32, no. 1, pp.89–102, October 2023.

[4] Kensuke Iizuka, Haruna Takagi, Aika Kamei, and Hideharu Amano, Power Analysis and Power Modeling of Directly-connected FPGA Clusters, *IEICE Transactions on Information and Systems*, vol. E106-D, no. 12, pp.1997–2005, December 2023.

International Conference Papers

[5] Kensuke Izuka, Haruna Takagi, Aika Kamei, Kazuei Hironaka, and Hideharu Amano, Power analysis of directly-connected FPGA clusters, In *Proc. of 2022 IEEE Symposium in Low-Power and High-Speed Chips (COOL CHIPS)*, pp.1–6, Tokyo, Japan, April 2022.

[6] Kimiyoshi Usami, Daiki Yokoyama, Aika Kamei, and Hideharu Amano, Optimal switching time to minimize store energy in MTJ-based flip-flops under process and temperature variations, In Proc. of 2022 IEEE Nordic Circuits and Systems Conference (NorCAS), pp.1–7, Oslo, Norway, October 2022.

Domestic Conference Papers and Technical Reports

[7] 亀井愛佳, 天野英晴, 小島拓也, 横山大輝, 宮内陽里, 宇佐美公良, 平賀啓三, 鈴木健太, 別所和宏, 不揮発性 FF を用いたマルチコンテキスト CGRA, 研究報告システムと LSI の設計技術 (SLDM), vol. 2021, no. 4, pp.1–6, 2021年12月.

[8] 亀井愛佳, 天野英晴, 小島拓也, 横山大輝, 宮内陽里, 宇佐美公良, 平賀啓三, 鈴木健太, 別所和宏, 不揮発性FFを用いたCGRA設計探索のためのばらつきを考慮したMTJへの書き込みエネルギー推定モデルの提案, 研究報告システム・アーキテクチャ (ARC), vol. 2022, no. 26, pp.1–7, 2022年3月.

[9] 亀井愛佳, 天野英晴, 小島拓也, 宇佐美公良, 平賀啓三, 鈴木健太, 別所和宏, 不揮発性FFのエネルギーモデルを用いた間欠動作アプリケーションのエネルギー最小化, DAシンポジウム2023論文集, pp.235–242, 2023年8月.

[10] 横山大輝, 宇佐美公良, 亀井愛佳, 天野英晴, MTJベース不揮発性フリップフロップの最適ストア時間に関する解析式の提案, 研究報告システムと LSI の設計技術 (SLDM), vol. 2022, no. 21, pp.1–6, 2022年11月.

[11] 飯塚健介, 高木春奈, 亀井愛佳, 弘中和衛, 天野英晴, FPGAクラスタのための電力測定ツールの導入と消費電力の分析, 信学技報, vol. 122, no. 60, pp.80–85, 2022年6月.

[12] 飯塚健介, 亀井愛佳, 弘中和衛, 天野英晴, FPGAクラスタの電力推定モデルの提案, 研究報告システムと LSI の設計技術 (SLDM), vol. 122, no. 451, pp.66–71, 2023年3月.

[13] 小島拓也, 亀井愛佳, 矢内洋祐, 天野英晴, 久我守弘, 飯田全広, Jupyter Notebook を介した RISC-V SoC 向け実機テスト環境の構築, 研究報告組込みシステム (EMB), vol. 2023, no. 24, pp.1–7, 2023年5月.

[14] 茅島秀人, 亀井愛佳, 小島拓也, 天野英晴, 誘導結合無線通信インタフェースにおけるバスアービトレーション方法の検討, 研究報告システム・アーキテクチャ (ARC), vol. 123, no. 145, pp. 55-60, 2023年8月