This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TM-RF: Aging-Aware Power-Efficient Register File Design for Modern Microprocessors

Na Gong, *Member, IEEE*, Jinhui Wang, *Member, IEEE*, Shixiong Jiang,
and Ramalingam Sridhar, *Senior Member, IEEE*

***Abstract*— Modern microprocessors employ register files (RFs) for performance enhancement and achieving instruction level parallelism simultaneously. However, RF incurs large power consumption owing to the highly frequent access. Meanwhile, as technology scales, bias temperature instability has become a major reliability concern for RF designers. This paper presents an aging-aware trimodal register file (TM-RF) design to enhance the power efficiency. As instructions pass through the pipeline, TM-RF places the bit-cells in different modes based on the register activity, thereby achieving significant power reduction. To meet design constraints of different applications, we present four schemes to implement the proposed design, providing design flexibility. Additionally, with device selection and worst case sizing methodology, we mitigate aging-effect-induced RF reliability degradation. Simulation results on SPEC 2000 benchmarks demonstrate that TM-RF achieves up to 81.4% power savings and 17% reliability improvement on average, with minimal impact on performance.**

***Index Terms*— Leakage current, low power, bias temperature instability (NBTI/PBTI), process variation, register file (RF).**

I. Introduction

AS A fundamental part in modern microprocessors, register file (RF) enhances the performance by shrinking the performance gap between microprocessor and memory systems, as well as increasing instruction level parallelism through implementing register renaming [1]–[3]. However, with aggressive technology scaling, power efficiency and reliability have become two main challenges to RF designers.

First, these microprocessors are capable of fetching, decoding, renaming multiple instructions per clock cycle, and on average every instruction requires three accesses (two reads and one write) to RFs. Such frequent access results in increased power consumption [4]. This situation is further exacerbated in simultaneous multithreading (SMT) processors, where the access frequency is increased by multiple threads. It has been reported that RFs consume 25%–37% of the total power in modern microprocessors [5].

Second, as process variation continues to pose challenge to the reliable operation of RF, bias temperature instability (NBTI/PBTI)-induced aging effect is emerging as another important lifetime reliability issue in RFs owing to the following three reasons: 1) the NBTI/PBTI effect increases exponentially with temperature and RF is a typical temperature hot spot in a processor; 2) the imbalanced data and the resulting signal probability of RFs lead to an increase in reliability degradation [6]; and 3) as RFs are accessed frequently, corrupted data in RFs can easily propagate to other parts of microprocessors. Researchers have shown that considerable amount of errors affecting a processor usually are from the RF [7]. Accordingly, designing RF with aging awareness has become a major requirement along with power consideration.

Many techniques have been presented in the literature for reducing the power consumption of RF [9]–[18]. For each technique, there is a tradeoff between power efficiency, performance, and implementation cost. Nevertheless, none of these techniques have accounted for the lifetime reliability issue to make RF tolerant to aging effect. On the other hand, previous NBTI/PBTI mitigation techniques [19]–[22] lead to considerable power consumption overhead, which significantly influence the power efficiency of RF.

In this paper, we propose trimodal RF, referred to as TM-RF, a novel circuit/architecture co-design that targets power savings and tolerance to variations and aging effect. A new bit-cell circuit is presented to provide three different modes, exploring power savings opportunities of RF as instructions pass through the pipeline. We had earlier presented the basic idea of TM-RF in [8] with some preliminary results. In this paper, we extend our original work and make the following additional contributions.

1) We propose a variation and aging-aware worst case sizing methodology to compensate for variations and aging-effect-induced RF reliability deterioration (Section IV-C).
2) We develop two circuit-level implementation schemes with different tradeoffs between power savings and wakeup performance, called lp-TM and aggr-TM. Additionally, we propose a finger-based flexible layout scheme to implement lp-TM and aggr-TM (Section IV-D).

Manuscript received June 23, 2013; revised February 11, 2014; accepted June 9, 2014. This work was supported in part by the North Dakota EPSCoR under Grant FAR0022051 in part by the National Natural Science Foundation of China under Grants 61204040 and 60976028; in part by the Ministry of Education of China, Ph.D. Programs Foundation, under Grant 20121103120018; and in part by the Plan Program of Beijing Education Science and Technology Committee under Grant JC002999201301.

N. Gong is with the Department of Electrical and Computer Engineering, North Dakota State University, Fargo, ND 58102 USA (e-mail: na.gong@ndsu.edu).

J. Wang is with the College of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100124, China (e-mail: wangjinhui@bjut.edu.cn).

S. Jiang and R. Sridhar are with the Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY 14260 USA (e-mail: shixiong@buffalo.edu; rsridhar@buffalo.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVLSI.2014.2334136

1063-8210 © 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

3) We present two architectural schemes to accomplish the proposed TM-RF strategy, enabling flexible design suitable for different applications (Section V).
4) A comprehensive suite of simulations on TM-RF is performed and the enriched results are discussed, including the power efficiency, reliability, temperature profile, area overhead, performance impact, and sensitivity to different RF configurations (details are shown in Sections VI-B and VI-C).
5) Furthermore, the comparison of TM-RF against other existing RFs is presented. Simulation results show that the proposed TM-RF design can achieve noticeable power reduction and reliability improvement with minimal area overhead and little performance degradation (Section VI-D).

The rest of this paper is organized as follows. In Section II, we provide a review of related low-power and reliable RF design techniques. In Section III, we present the proposed TM-RF. In Sections IV and V, we discuss circuit implementations and architectural mechanisms of TM-RF, respectively. Section VI describes the simulation methodology and results, followed by the conclusion in Section VII.

II. Related Work

Significant amount of research that targets low-power and reliable RFs has been reported in the literature. In this section, we briefly review some existing work related to the proposed technique.

A. Related Work on Low-Power RF Design

Low-power RF design techniques can be broadly classified into three different categries.

1) Reducing RF Pressure: The first set of technqiues attemp to reduce RF pressure, and thus smaller RFs can be used to reduce the power consumption. Several techniques have been proposed in literature to achieve early register release including hardware-based scheme [9] and compiler-assist scheme [10]. These schemes can be used in conjunction with the design proposed in this paper. Another approach to reduce RF pressure is late register allocation [11]. In addtion, double-pumped RF [2] and banked RF [12] have been investigated to reduced the number of registers (ports) and accordingly the power consumption.

2) Reducing RF Access Frequency: In [13], several compiler assisted instruction scheduling techniques are developed and therefore the operand values are transferred via the bypass path instead of accessing RF. Balkan *et al.* [14] proposed a selective writeback technique which avoids the writeback of transient values into the RF, thereby reducing RF power consumption.

3) Alternative RF Structure: The third set of solutions modifies the RF structure mainly by exploring the power characteristics and the inefficiency in RF usage. Shioya *et al.* [15] observed that most destination results are short-lived operants and therefore they presented a caching approach to store the short-lived operants to a small RF. Ergin *et al.* [16] implemented checkpointed RF (CP-RF) to release registers early. This technique relies on a new checkpointed bit-cell which is capable of saving register copies and enables recovery from branch misprediction. With the fact that storing 0 is more power efficient than storing 1, Jin *et al.* [17] designed discharge-RF, which enables bit-cell discharging once the stored information is invalid. More recently, Shieh and Chen [18] proposed a monitoring RF scheme (monitor-RF) which uses reorder buffer (ROB) to monitor each incoming instruction and informs DVS controller to adjust the voltage levels of its destination register, achiving power reduction.

Our paper falls into the third category. Among the aforementioned related works, discharge-RF, CP-RF, and monitor-RF are close to our design. Compared with these techniques, the proposed TM-RF is more effective in power reduction with reliability improvement. Section VI-D provides more details comparing the proposed technique with these related works.

B. Related Work on Reliable RF Design

Past works on NBTI/PBTI mitigation mostly either attemp to reduce the imbalance rate of RF or balancing the device degradation in bit-cells.

In [19], periodic register rotation was presented to reduce the mismatch between SRAM cell inverter pairs induced by NBTI. In [20], an adaptive body bias technique was introduced to reduce the impact of NBTI and process variation. Most recently, Siddiqua and Gurumurthi [21] proposed recovery boosting RF which allows both pMOS devices in the bit-cell to be put into the recovery mode. In [22], a hybrid-cell RF design was presented to mitigate NBTI-induced degradation by storing more vulnerable data bits in the robust 8T cells and the less vulnerable bits in the conventional 6T cells.

The common feature of these techniques is that the reliability improvement comes at a cost of increased power consumption. In contrast, our technique realizes reliability improvement and significant power savings simultaneously.

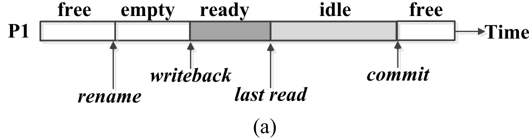
III. Overview of the Proposed Technique

In this section, we first present the motivation of the proposed technique that explores power savings opportunities as instructions pass through the pipeline. Then, the high-level overview of the proposed TM-RF is shown.

A. Motivational Example

As mentioned previously, modern microprocessors adopt register rename technique to enhance instruction-level parallelism. This technique maps architectural registers to physical registers and uses a register alias table (RAT) to keep track of the state of each physical register [1]. A physical register has four states according to the usage of its content: empty, ready, idle, and free [9], as shown in Fig. 1(a). Consider the physical register P1 in the code example shown in Fig. 1(b). When instruction A is renamed, P1 is mapped to architectural register r1. Accordingly, P1 changes from free state to the empty state in RAT; as A executes, P1 becomes the ready state to contain the valid data; after its last consumer

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



(a)

PC	Original code	Rename code	ROB
--	-----------	-------	-
A	ADD -> r1	ADD ->P1	P1
B	OR <- r1	OR<- P1	

Many instructions do not need r1
Worst scenario: long latency events happen such as L2 cache misses

PC	Original code	Rename code	ROB
L	SUB ->r1	SUB ->P2	P2
M	BRANCH		-
N	ADD-> r3	ADD P9	P9
O	BRANCH		
--	----------	------	

(b)

States	empty	ready	idle	free
Time (%)	22.6	20.0	18.3	39.1

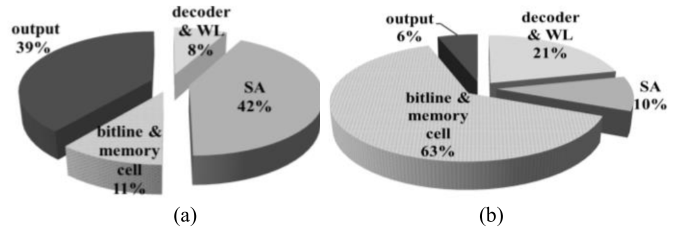
(c)

Fig. 1. (a) Different states of registers. (b) Example code sequence. (c) Register state time distribution.

(LastUser) instruction B reads its value, P1 is in idle state for recovery until the instruction L redefining its architectural register (Redefiner) commits; and then P1 returns to the free state. In the worst case scenario, long latency events such as L2 cache misses occur between the LastUser (instruction B) and the Redefiner (instruction L). However, L2 cache misses may take hundreds of cycles to resolve [23]. During such long service time, P1 stays in the idle state, consuming large leakage power.

Therefore, the motivation for our paper arises from the following two observations.

1) As instructions pass through the pipeline, registers only spend a small fraction of its life time in ready state and therefore there is a large opportunity to reduce RF power consumption. Fig. 1(c) shows the simulation results of average register state distribution on the integer SPEC2000 benchmarks. (The experimental configuration will be presented in Section VI-A.) It is shown that the average ready time of registers is only 18.3%. Such small ready time is due to two reasons: 1) most registers are read at most once; and 2) some registers are never read because their values are not needed or their consumers obtain the result through bypass logic [12].
2) The contribution of RF bit-cells to leakage power is much larger than that to dynamic power. Using a modified version of CACTI 5.3 [24], we model a 128-entry $\times$ 64 b two-read/one-write ported RF and the breakdown of dynamic power and leakage power of RF components is shown in Fig. 2. It can be seen that bit-lines and bit-cells only consume 11% dynamic power, but they contribute to 63% leakage power. Because most of the leakage current of bit-lines flows from the bit-cells, reducing the leakage in bit-cells can eliminate the bit-line leakage [25]. Accordingly, we can suppress the leakage power of the entire RF effectively with low leakage RF bit-cells.



(a) (b)

Fig. 2. (a) Dynamic power and (b) leakage power breakdown of 128-entry $\times$64 b two-read/one-write ported RF.

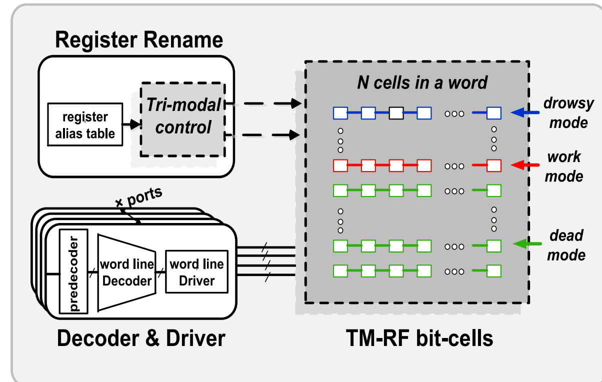


Fig. 3. Overview of the proposed TM-RF.

B. Overview of the Proposed TM-RF

Fig. 3 shows the proposed design. The TM-RF bit-cell provides three different modes: normal work mode, low-leakage data-retention drowsy mode, and minimum-leakage dead mode. The fundamental idea is that, based on the states of a register, the trimodal control logic generates control signals and then passed them to RF, placing the corresponding bit-cells into the appropriate mode. Specifically, when a register is in ready state, its bit-cells are in work mode to keep valid data effectively; as the register enters the idle state, its bit-cells can be placed in drowsy state for recovery; once the register is released to free state, its bit-cells become dead mode without keeping valid information. The circuit-level implementations and architectural mechanisms of TM-RF will be discussed in Sections IV and V, respectively.

IV. TM-RF: Circuit Implementation

In this section, we discuss the circuit implementations of the proposed TM-RF. For accurate characteristics, we use 32-nm PTM High k/Metal-gate technology [26] (low-$V_{\mathrm{th0_nMOS}}$ = |low-$V_{\mathrm{th0_pMOS}}$| = 0.49 V; high-$V_{\mathrm{th0_nMOS}}$ = |high-$V_{\mathrm{th0_pMOS}}$| = 0.65 V; V_{DD} = 0.9 V) and perform HSPICE simulations in the following analysis. To capture the process variation, the threshold voltage (V_{th}) variation for an nMOS (pMOS) transistor with minimum size is 24 mV (29.2 mV) [27].

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

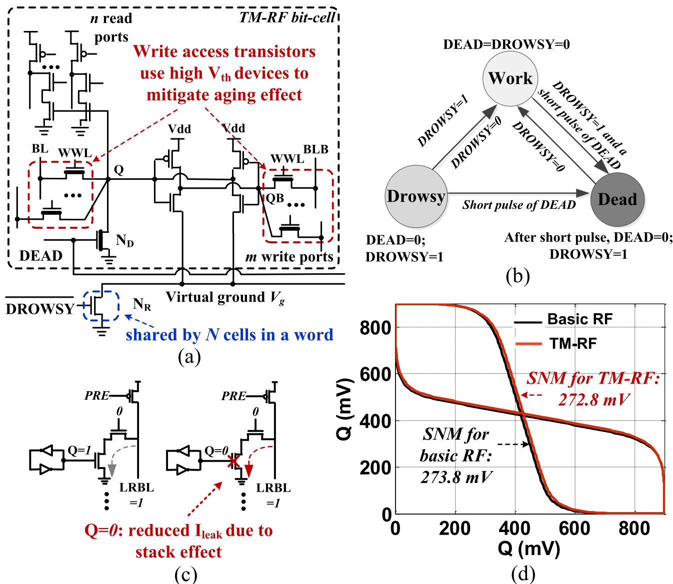


Fig. 4. (a) Proposed trimodal RF bit-cells. (b) Transition diagram. 1 and 0 represent logic high and low, respectively. (c) Data-dependent leakage characteristics of RF bit-cells. (d) SNM of TM-RF.

A. TM-RF Bit-Cells

Fig. 4(a) shows the schematic of TM-RF bit-cell and Fig. 4(b) shows its transition diagram. As shown, a discharging nMOS transistor (N_D) is inserted to each conventional RF bit-cell; and a data-retention nMOS transistor (N_R) is connected to each register. Accordingly, each bit-cell has three different modes determined by two control signals: DEAD and DROWSY.

1) Normal Work Mode (DEAD = DROWSY = 0): In a TM-RF bit-cell shown in Fig. 4(a), if DEAD=DROWSY = 0, the inserted N_D is maintained cutoff and N_R is turned on. Therefore, these additional devices do not interfere with normal operations on bit-cells and the working function of TM-RF is similar to a conventional one.

2) Low-Leakage Data-Retention Drowsy Mode (DEAD = 0 and DROWSY = 1): To realize drowsy mode, the gated-ground technique [28] is used in our design. While N_D is turned off, because of the charging leakage current, the node storing 0 goes up and gets saturated quickly. This saturated voltage depends on the size and V_{th} of N_R. By selecting the appropriate size of N_R, the nMOS transistor which connects the node storing 1 to virtual ground V_g. Therefore, the node storing 1 is firmly strapped to V_{dd} and the data is retained successfully. Because of the stack effect, the leakage current of a register is suppressed effectively. Note that, the size of N_D is an important design consideration to achieve a successful data retention, which will be discussed in Section IV-C.

3) Minimum-Leakage Dead Mode (Short DEAD Pulse and DROWSY = 1): As shown in Fig. 4(c), the bit-line leakage current of all read ports strongly depends on the stored data in bit-cells: the leakage current of storing 0 is much smaller than that of storing 1 due to the stack effect. Therefore, similar to Discharge-RF in [17], we insert a high V_{th} device N_D to each bit-cell and N_D is controlled by the signal DEAD. If DEAD is enabled, the data stored in a bit-cell will be discharged, minimizing the leakage current. Unlike [17], which adopts a dc DEAD signal, we apply a short-pulse DEAD signal: when a register is released, the DEAD signal to its bit-cells becomes high to discharge N_D; once this process is finished, DEAD is changed to zero immediately to turn off N_D. Our experiment in Section IV-C shows that this discharging process is very short (~18 ps) compared with the clock period (193 ps). Such short-pulse DEAD signal results in two advantages: 1) it achieves gate leakage current savings, because the reverse gate leakage current in OFF N_D is much less than the forward gate leakage current in ON N_D [36]; and 2) the DEAD pulse turns off N_D very early, and therefore there is no need to set this signal when the bit-cells return the work mode, achieving a fast and robust transition. The generation logic of DEAD signal will be discussed in Section V.

The dead mode of TM-RF bit-cells provides two additional advantages in terms of power efficiency.

1) When a physical register has been allocated to an architectural register, but before the valid data is written, the register is in empty state and its bit-cells are placed in minimum-leakage dead mode automatically, achieving high power efficiency. From power consumption point of view, TM-RF achieves the same effect as late register allocation technique [11].
2) Similar to the Discharge-RF in [17], because a register with the proposed design stores a 0 by default, writing a 0 to an empty register can be avoided. This can be done by the zero detection as implemented in [17] with negligible hardware overhead. Accordingly, TM-RF can result in further power savings, which will be discussed in Section VI-B.

Note that, the additional asymmetry of TM-RF caused by N_D has a negligible impact of static noise margin (SNM) of RF during the read operation owing to the decoupled read path [as shown in Fig. 4(d)]. However, as more 0s are stored in RF in the dead mode, TM-RF may induce larger signal probability imbalance, increasing the impact of aging effect. Accordingly, we introduce the aging effect awareness to TM-RF to mitigate the impact of NBTI/PBTI.

B. Aging-Effect-Aware TM-RF Design

1) Aging Effect on RF Reliability: The reliability degradation due to NBTI/PBTI has become a major concern in RF design with the introduction of high-k and metal-gate process (HK + MG) in 32-nm and smaller technologies [6], [29]. NBTI/PBTI occurs as a negative/positive gate to source voltage is applied to pMOS/nMOS, generating interface traps and resulting in V_{th} shift. With reaction–diffusion (R–D) mechanism, V_{th} increase owing to the long-term NBTI/PBTI effect can be expressed as [6]

$$\Delta V_{\rm th} = K_{AC} \cdot t^n = \alpha\,(S) \cdot K_{\rm DC} \cdot t^n \tag{1}$$

$$K_{\rm DC} = A \cdot T_{\rm ox} \cdot \sqrt{C_{\rm ox}\,(V_{\rm GS} - V_{\rm th})} \cdot \exp\left(\frac{E_{\rm ox}}{E_0}\right) \cdot \left[1 - \frac{V_{\rm ds}}{\alpha\,(V_{\rm GS} - V_{\rm th})}\right] \cdot \exp\left(-\frac{E_a}{kT}\right) \tag{2}$$

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

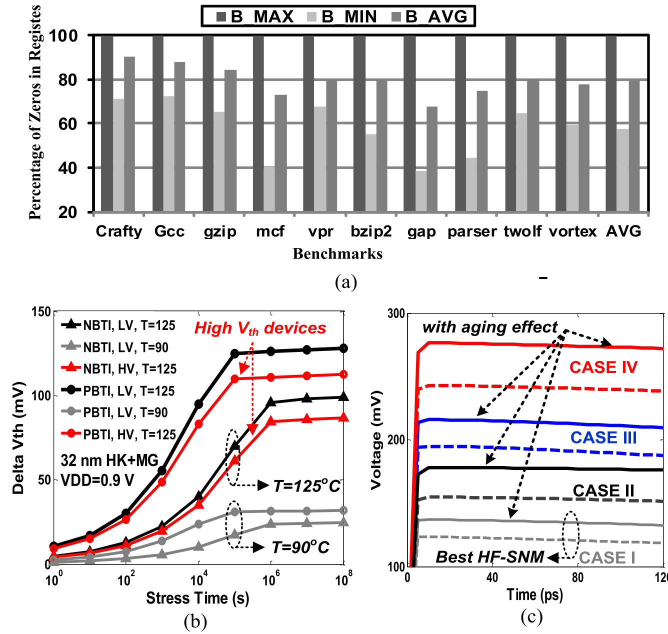


Fig. 5. (a) Probability of 0 in RF bit-cells for different applications. B_MAX: maximum value across different bit-cells in a register. B_MIN: minimum value across different bits in a register. B_AVG: average value across different bits in a register. (b) Dependency of aging-effect-induced V_{th} shift on temperature and V_{th}. (c) Write half-select disturb in four different cases.

where k is the Boltzmann constant, C_{ox} is the oxide capacitance per unit area, T_{ox} is the gate oxide thickness, T is the temperature, A, E_0, E_{a}, α, and δ_{v} are constants equal to 1.8 mV/nm/C$^{0.5}$, 2.0 MV/cm, 0.13 eV, 1.3, and 5.0 mV, respectively. In addition, the ac factor $\alpha(S)$ is a function of stress probability (S). Note that the ac factor also depends on the duty frequency [38]–[41]. Here we neglect the impact of duty frequency, since it is relatively insignificant as compared with the impact of stress probability [29], [42]. For devices in RF bit-cells, $\alpha(S)$ is determined by the stored data probability of bit-cells. The best condition is the stored 0 probability equals to 1 probability, which induces the balance V_{th} increase in two bit-cell inverters and minimizes the aging effect. We extract the stored 0 probability of RF bit-cells shown in Fig. 5(a) based on integer SPEC 2000 benchmarks (the experimental configuration will be presented in Section VI-A). We can see that, the average of 0 probability in RF bit-cells (B_AVG) is as high as 79.23%, resulting in significant imbalanced stress probability for devices connected to Q and QB in RF bit-cells [see Fig. 4(a)]. Such imbalanced V_{th} degradation in two bit-cell inverters leads to significant aging effect on RF. In our analysis, we account for the impact of data probability on aging effect and incorporate the V_{th} drift into our simulation, which we will detail in Section VI-A.

We calibrate K_{DC} to match the hardware data results in [30], thereby fitting the saturation of trap density on HK + MG interface under long-term stress. Accordingly, we calculate the V_{th} shift caused by NBTI effect after seven years, which is the typical lifetime of modern microprocessors [31]. Fig. 5(b) shows the V_{th} drift induced by dc NBTI/PBTI stress. As observed, the V_{th} shift increases significantly with temperature. Another important observation is that devices with higher V_{th} are less sensitive to the aging effect. This is due to the dependence of K_{DC} on temperature and V_{th} as given in (2).

2) RF Design With Aging Awareness: The aging-effect-induced V_{th} increase influences the performance of SRAM cells including read stability, write margin, access time, and leakage power. Recent work [29] has shown that, the half-select SNM (HS-SNM) dominates the stability of 8 T bit-cells, and therefore we use HS-SNM as the lifetime reliability metric in our analysis. As defined in [29], HS-SNM is the voltage difference between the disturb voltage on storage nodes and the trip voltage of the inverters (V_{trip}) in half-select bit-cells during write operations.

Since high V_{th} devices help mitigate NBTI/PBTI effect, we use high V_{th} devices in TM-RF bit-cell design. To keep read access speed, the two transistors in the read stack of each read-port adopts low V_{th} devices. To improve reliability, we consider four different cases.

1) CASE I: only two write access nMOS transistors use high V_{th} devices.
2) CASE II: all write access nMOS transistors and two bit-cell inverters are low V_{th} devices.
3) CASE III: all write access nMOS transistors and two bit-cell inverters use high V_{th} devices.
4) CASE IV: only the transistors in two bit-cell inverters use high V_{th} devices.

Fig. 5(c) shows the write half-select disturb in four cases and the solid curves indicate the results considering aging effect. As can be seen, CASE I with only two high V_{th} write access transistors is able to improve the HS-SNM compared with other conditions. Accordingly, the write access transistors in TM-RF bit-cell adopt high V_{th} devices [as shown in Fig. 4(a)]. Note that, the high V_{th} write access transistors result in slight longer write delay. However, the performance of RF is typically limited by the read access time, so high V_{th} write access transistors induce little performance degradation, which will be discussed in Section IV-E.

C. Implementation Considerations

To lower the layout area overhead of TM-RF, we use minimum-sized transistor for N_{D}: $W_{\text{ND}} = W_{\text{min}}$ and $L_{\text{ND}} = L_{\text{min}}$. Note that Wmin is the minimum width allowed in the technology and it varies for different technologies. Here, for the 32-nm technology we use, we assume $W_{\text{min}} = L_{\text{min}} = 32$ nm.

Moreover, in the proposed scheme, N_{R} requires a careful sizing process because it determines the value of V_{g}, as discussed in Section IV-A. A larger V_{g} can result in more power savings, but it is easier to cause data-retention failure. Thus, sizing N_{R} to account for the aging effect and process variation is required to prevent potential data-retention failure in late chip lifetime. In addition, the value of V_{g} influences wakeup delay when a bit-cell enters work mode. Furthermore, to avoid RF access penalty, we need to make sure the write operation can be finished in one cycle. Therefore, to compensate for variations and aging-effect-induced RF reliability deterioration, we propose a variation and aging-aware worst case sizing

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

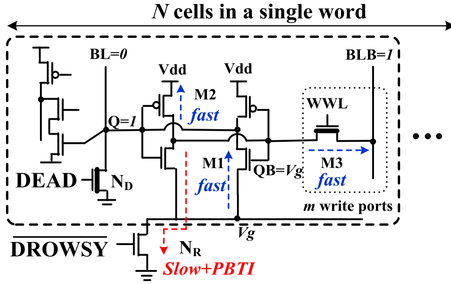


Fig. 6. Process variation and aging-effect-aware worst case analysis for data retention.

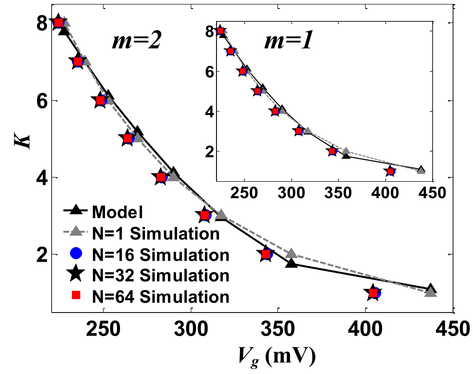


Fig. 7. V_g and K.

methodology while considering three factors: 1) data-retention stability; 2) wakeup delay time; and 3) write time.

1) Data-Retention Stability: Fig. 6 shows the process variation and aging-effect-induced worst case for data retention in a TM-RF bit-cell. As shown, in the worst condition, TM-RF generates maximum V_g. Accordingly, the large subthreshold leakage current (I_{sub}) of fast M1, M2, and M3 flow out of V_g and into V_{dd}, while small I_{sub} of slow N_R considering PBTI effect flows into ground. Equating the currents at the V_g node shown in Fig. 6, we get

$$I_{sub}(N_R) = I_{sub}(M1) + I_{sub}(M2) + m \cdot I_{sub}(M3) \quad (3)$$

where m is the number of write ports of RF. According to BSIM4 model, I_{sub} is given by

$$I_{sub} = A \cdot \left(\frac{W}{L}\right) \cdot \exp\left[\frac{q}{nKT}\left(V_{GS} - V_{th0} - \gamma' V_{SB} + \eta V_{ds}\right)\right] \cdot \left(1 - e^{-(qV_{ds}/kT)}\right)$$
$$A = \mu_0 C_{ox}\left(\frac{kT}{q}\right)^2 e^{1.8} \quad (4)$$

where γ' is the linearized body-effect coefficient, η is the DIBL coefficient, n is the subthreshold swing coefficient, C_{ox} is the gate-oxide capacitance, μ_0 is the zero-bias mobility.

Now substituting (4) in (3), we get

$$\left(\frac{W}{L}\right)_{NR} \cdot \exp\left[\frac{q}{nKT}\left(-V_{tnlow}^{slow+PBTI} + \eta V_g\right)\right] = \sum_{i=1}^{N}\left\{\left(\frac{W}{L}\right)_{M1} \cdot \exp\left[\frac{q}{nKT}\left(-V_{tnlow}^{fast} - \gamma' \cdot V_g + \eta \cdot (V_{dd} - V_g)\right)\right] + \left(\frac{W}{L}\right)_{M2} \cdot \exp\left[\frac{q}{nKT}\left(V_{tplow}^{fast} + \eta \cdot (V_{dd} - V_g)\right)\right] + m \cdot \left(\frac{W}{L}\right)_{M3} \cdot \exp\left[\frac{q}{nKT}\left(-V_g - V_{tnhigh}^{fast} - \gamma' \cdot V_g + \eta \cdot (V_{dd} - V_g)\right)\right]\right\} \quad (5)$$

where $V_{tnlow}^{slow+PBTI}$, V_{tnlow}^{fast}, V_{tplow}^{fast}, V_{tnhigh}^{fast} and are V_{th} of N_R, M1, M2, and M3, respectively, in worst variation and aging effect condition (see Fig. 6).

We denote $(W/L)_M = R_M (W/L)_{min}$ and (5) becomes

$$\frac{1}{N} \cdot R_{NR} = R_{M1} \cdot C_1 \cdot \alpha_1^{V_g} + R_{M2} \cdot C_2 \cdot \alpha_2^{V_g} + m \cdot R_{M3} \cdot C_3 \cdot \alpha_3^{V_g} \quad (6)$$

where

$$C_1 = \exp\left[\frac{q}{nKT}\left(V_{tnlow}^{slow+PBTI} - V_{tnlow}^{fast} + \eta \cdot V_{dd}\right)\right]$$
$$C_2 = \exp\left[\frac{q}{nKT}\left(V_{tnlow}^{slow+PBTI} + V_{tplow}^{fast} + \eta \cdot V_{dd}\right)\right]$$
$$C_3 = \exp\left[\frac{q}{nKT}\left(V_{tnlow}^{slow+PBTI} - V_{tnhigh}^{fast} + \eta \cdot V_{dd}\right)\right]$$
$$\alpha_2 = \exp\left[\frac{q}{nKT}(-2\eta)\right] \alpha_1 = \exp\left[\frac{q}{nKT}\left(-\gamma' - 2\eta\right)\right]$$
$$\alpha_3 = \exp\left[\frac{q}{nKT}\left(-\gamma' - 2\eta - 1\right)\right] \quad (7)$$

where C_1–C_3 and α_1–α_3 are technology and design-dependent constants.

To capture the effect of N_R size and the number of bits N, we define a parameter donated by K, which is the ratio of R_{NR} to N. Therefore, (6) becomes

$$K = R_{M1} \cdot C_1 \cdot \alpha_1^{V_g} + R_{M2} \cdot C_2 \cdot \alpha_2^{V_g} + m \cdot R_{M3} \cdot C_3 \cdot \alpha_3^{V_g}. \quad (8)$$

With BSIM4 parameters, $0 < \alpha_1, \alpha_2, \alpha_3 \ll 1$. Therefore, the expotential terms can accurately be approximated using their second-order Taylor series expanision. Hence, (8) is simplified to

$$K = k_1 \cdot V_g^2 + k_2 \cdot V_g + k_3 \quad (9)$$

where

$$k_1 = \frac{R_{M1} \cdot C_1 \cdot (\ln \alpha_1)^2 + R_{M2} \cdot C_2 \cdot (\ln \alpha_2)^2 + m \cdot R_{M3} \cdot C_3 \cdot (\ln \alpha_3)^2}{2}$$
$$k_2 = R_{M1} \cdot C_1 \cdot \ln \alpha_1 + R_{M2} \cdot C_2 \cdot \ln \alpha_2 + m \cdot R_{M3} \cdot C_3 \cdot \ln \alpha_3$$
$$k_3 = R_{M1} \cdot C_1 + R_{M2} \cdot C_2 + m \cdot R_{M3} \cdot C_3. \quad (10)$$

Therefore, (10) captures the approximately square relationship between K and V_g. Fig. 7 compares the derived equations against SPICE simulations, demonstrating acceptable accuracy.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

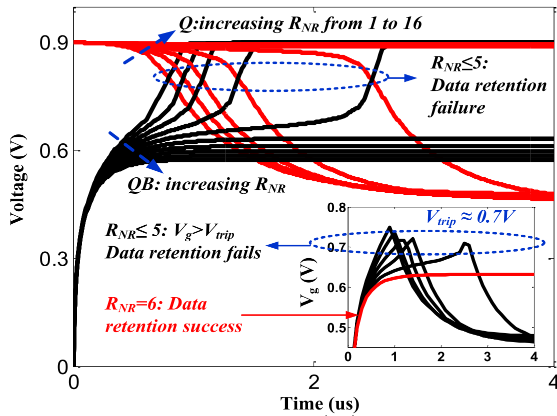


Fig. 8. Impact of $R_{\rm NR}$ on data-retention ability of 64-bit RF.

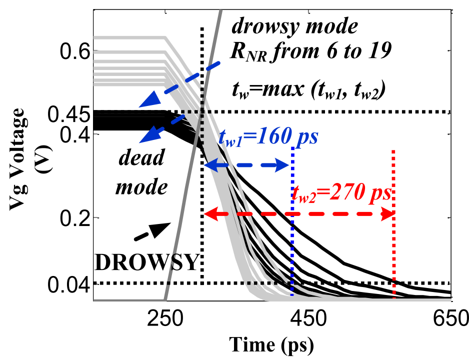


Fig. 9. t_w and the size of N_R at the presence of process variation and aging effect.

To achieve successful data retention, V_g is required to be smaller than $V_{\rm trip}$ ($V_{\rm g} < V_{\rm trip}$), so we get

$$R_{NR} \geq \frac{N}{k_1 \cdot V_{\rm trip}^2 + k_2 \cdot V_{\rm trip} + k_3}. \quad (11)$$

From (11), it can be seen that the number of bits N influences $R_{\rm NR}$ significantly. As N increases, $R_{\rm NR}$ has to be increased to provide the data-retention ability. We perform SPICE simulations on 64-bit TM-RF as $R_{\rm NR}$ varies and the result is shown in Fig. 8. As observed, to achieve successful data-retention, the minimum $R_{\rm NR}$ is 6.

2) Wakeup Delay: The second design concern is the wakeup delay of TM-RF. As shown in Fig. 9, defined for a RF operating in the dead or drowsy modes, the wakeup delay (t_w) measures the maximum delay between the time when the DROWSY signal crosses the 50% $V_{\rm dd}$ level as it makes a transition to the work mode and the time when the V_g node reaches 5% of the $V_{\rm dd}$ level as it is discharged to zero.

The effect of N_R size on the wakeup delay is similar to the traditional power gating analysis [32]. This wakeup delay is expressed as

$$t_w = \frac{\int I_{\rm NR}(t) \cdot dt}{I_{\rm ON,NR}} \cong \frac{V_g \cdot C_{\rm word}}{I_{\rm ON,NR}} \quad (12)$$

where $I_{\rm NR}$ is the current of N_R after it turned on to wake up N bit-cells in a register, $I_{\rm ON,NR}$ is the current of N_R and $C_{\rm word}$ is the total capacitance of N bit-cells. As increasing $R_{\rm NR}$ reduces V_g, according to (12), such reduction lowers the delay.

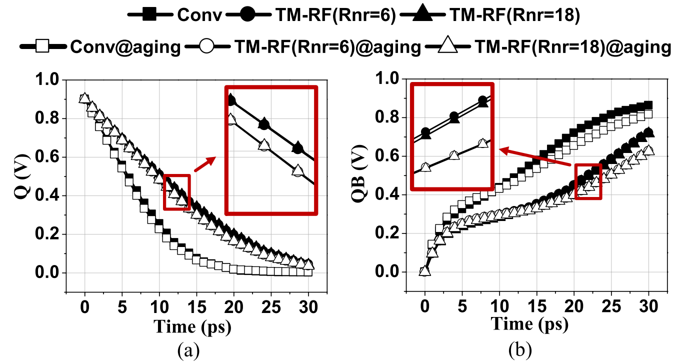


Fig. 10. Write time of TM-RF with/without aging effect. (a) Q. (b) QB.

Fig. 9 shows t_w of TM-RF with different $R_{\rm NR}$ under process variation and aging effect. As shown, as $R_{\rm NR}$ is 6, the wakeup delay is two cycles; to keep the wakeup delay within a single cycle, we need $R_{\rm NR} \geq 18$.

3) Write Time: The third design concern is to ensure that the write operation of TM-RF can be finished in one cycle, and therefore the access time of RF is still determined by the read access time. As discussed in Section IV-B, the high $V_{\rm th}$ write access transistors in TM-RF bit-cells induce increased write time. Also, the write time is influenced by the size of N_R.

We evaluate the write time of TM-RF as $R_{\rm NR}$ are 6 and 18 with and without considering aging effect and the results are shown in Fig. 10. The write time is slightly longer under aging effect for both conventional RF and TM-RF. Our simulation shows that the read access time of 128-entry ×64 b four-read two-write ported TM-RF without aging effect (with aging effect) is 193 ps (194 ps), achieving ∼5 GHz operating frequency. Therefore, though the write time of TM-RF is increased, the write operation still can be finished during one cycle, as shown in Fig. 10. Note that, the proposed TM-RF increases the read access time from 190 to 193 ps without considering aging effect, so there is a very slight increase in the RF read access time owing to the longer word select and bit lines. At the presence of seven-year aging effect, the read access time of TM-RF is similar to that of conventional RF (∼194 ps) because of its effectiveness in mitigating the NBTI/PBTI effect. As a result, the proposed TM-RF shows negligible effect on read access time (<2%).

4) Statistical Evaluation: To further illustrate the effectiveness of the adopted worst case sizing methodology in the presence of process variation, voltage variation, and aging effect, we run Monte Carlo (MC) simulations to obtain a statistical measure. Instead of worst process variation and aging effect condition, we consider 3σ process variation and aging-effect-induced $V_{\rm th}$ shift in all devices. Furthermore, the supply voltage is assumed to have an independent normal Gaussian distribution with 3σ variation of 10%, according to the International Technology Roadmap for Semiconductors [37]. One thousand Monte Carlo simulations are done to achieve enough statistical accuracy.

Fig. 11 shows the statistical results for 1000 samples. In Fig. 11(a), we present the distribution of data-retention

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

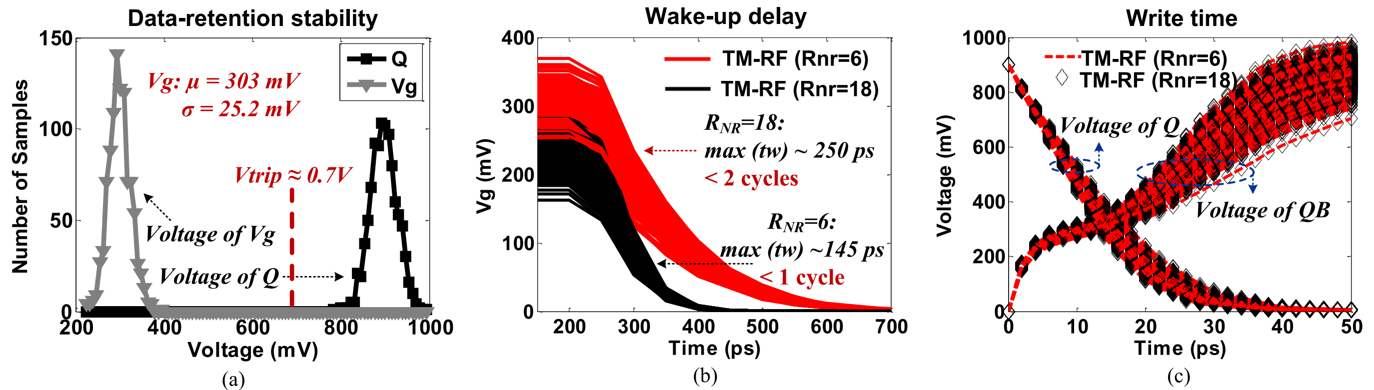


Fig. 11. Statistical evaluation results based on Monte Carlo Simulations. (a) Data-retention stability. (b) Wakeup delay. (c) Write time.

stability. Although V_g is affected considerably by the variations and aging effect during data-retention process ($\sigma = 25.2$ mV), it is consistently smaller than V_{trip} and therefore the data can be stored successfully. As shown in Fig. 11(b), the number of cycles needed for waking up from the drowsy/sleep mode are 1 and 2 for $R_{\text{NR}} = 18$ and $R_{\text{NR}} = 6$, respectively. In terms of write operation as shown in Fig. 11(c), it can be completed during one cycle for all samples. Therefore, the adopted worst case sizing methodology is very tolerant to variations and aging effect.

From the above analysis, we can see that sizing N_R is a tradeoff between power and wakeup delay, which is required to meet the design constraints. The constraints and the priority of the parameters are application specific. In high-performance applications such as leading servers are performance driven with large power budget, most design decisions are made to deliver a guaranteed performance. Alternatively, in embedded systems such as the multimedia applications, RFs consumes about 50% data-path power consumption though the performance can be compromised. Thus, the design effort mainly focuses on power savings.

Here, we develop two different schemes to meet the variation of power-performance requirements: 1) low-power TM-RF (lp-TM) with $R_{\text{NR}} = 18$; and 2) aggressive TM-RF (aggr-TM) with $R_{\text{NR}} = 6$. lp-TM has low wakeup delay (one cycle) and less power savings, while aggr-TM yields a larger power reduction at a higher wakeup delay cost (two cycles).

D. Layout Implementations of TM-RF Bit-Cell

With the MOSIS deep submicrometer design rules [33], we implemented the layout design of a four-read/two-write ported TM-RF bit-cell and it is shown in Fig. 12(a). In a typical SRAM array as shown in Fig. 12(b), the cells are laid out a mirrored fashion and therefore the same interconnect can be shared by adjacent cells. TM-RF bit-cell has a similar topology to that of the conventional bit-cell except the routing of N_D and N_R. We implement N_R as a column of transistors placed at one side of the memory cell array. In addition, to achieve two schemes, we propose a finger-based flexible layout scheme [as shown in Fig. 12(c)].

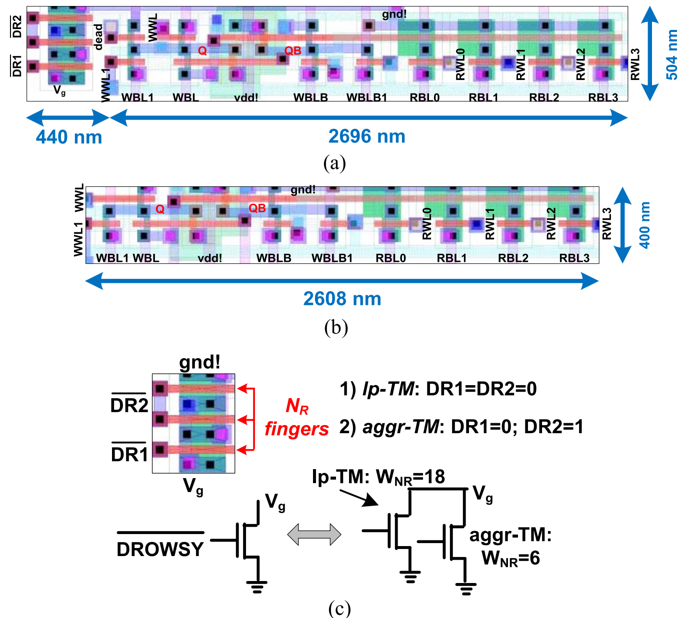


Fig. 12. Layout implementation. (a) Four-read/two-write ported TM-RF bit-cell. (b) Conventional four-read/two-write ported RF bit-cell. (c) Finger-based flexible design.

The lp-TM and aggr-TM schemes are provided by controlling the fingers of N_R devices with two control signals (DR1 and DR2):

1) *lp-TM*: DR1 = DR2 = 0, N_R fingers are completely turned on and $R_{\text{NR}} = 18$;
2) *aggr-TM*: DR1 = 0 and DR2 = 1, only the bottom finger is turned on and $R_{\text{NR}} = 6$.

As shown in Fig. 12, four-read/two-write ported TM-RF bit-cell consumes 51% more area than the conventional design. This is because, V_g cannot be shared between two adjacent rows, the new design results in a taller cell layout. At the same time, the additional N_R and N_D lead to a wider cell layout as compared with the conventional design. It should be noted that, similar to the bit-cell of CP-RF [16], this area overhead is not proportional to the number of RF ports. Therefore, adding N_D and N_R to a heavily multiported RF is expected to induce only a small amount of area overhead. For example, for a

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE I
POWER CONSUMPTION OF 64-b RFs $T = 90$ °C

Power	RFs		'0'	'1'
Read (μW)	Basic		0.76	119.1
	TM-RF (work)	aggr-TM	0.75	144.1
		lp-TM	0.75	144.1
Write(μW)	Basic		270	275
	TM-RF (work)	aggr-TM	283.1	291.3
		lp-TM	289.6	291.2
Leakage (nW)	Basic		421.1	4175
	TM-RF (drowsy)	aggr-TM	9.675	37.05
		lp-TM	16.27	41.34
	TM-RF (dead)	aggr-TM	9.675	-
		lp-TM	16.27	-

12-read/6-write ported TM-RF, its bit-cell area overhead can be reduced to 30%. Given the power savings discussed as follows, this area increase is tolerable.

E. Circuit-Level Evaluations on Power Efficiency

Table I presents the power consumption comparison of a 64-bit four-read/two-write TM-RF and a conventional one. In our simulation, the temperature is 90 °C, which is the average operating temperature for high-performance processors [21]. As expected, because of the asymmetrical structure of RF bit-cells, the power characteristics are data dependent. In particular, operations with 0 consistently consume less power than that with 1.

As observed in Table I, the read and write power of TM-RF in work mode are increased by the inserted transistors. An interesting observation is that the read power of TM-RF storing 0 is even reduced as compared with the conventional one. This is because the high V_{th} write access transistors in TM-RF suppress the active leakage current effectively.

For leakage power in drowsy and dead modes, the power savings of TM-RF are significant. The leakage power of a basic register with 1 and 0 are 4175 and 421.1 nW, respectively. As compared, if the proposed register with 1 is in drowsy mode, it consumes 37.05 nW (aggr-TM) and 41.34 nW (lp-TM), respectively. The drowsy mode containing 0 consumes the same power as the dead mode and it only consumes 9.675 nW (aggr-TM) and 16.27 nW (lp-TM), respectively, achieving up to 99.77% power savings. Note that, TM-RF can yield further power savings by avoiding writing 0 to an empty register, which will be discussed in Section VI-B.

V. TM-RF: ARCHITECTURAL CONTROL IMPLEMENTATION

In this section, the architectural mechanisms of the proposed TM-RF are explored in modern microprocessors.

The TM control logic is set by the rename logic, as the register renaming logic tracks the state of physical registers. Because most renaming logic is placed close to RF such as Alpha 21 264, the power consumed by the signal transmission from renaming logic to RF can be negligible. Here, we develop two schemes with different implementation complexity.

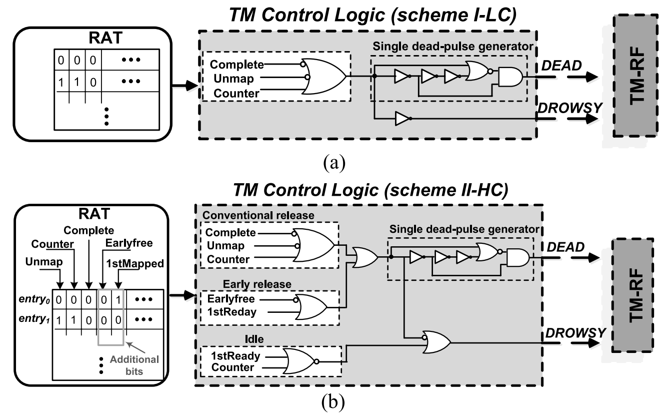


Fig. 13. TM control logic. (a) Scheme I-LC. (b) Scheme II-HC.

A. Scheme I-Low Complexity

We first present a simple scheme with minimal hardware support to implement TM-RF [as shown in Fig. 13(a)]. In a conventional superscalar microprocessor, the renaming logic has one entry for each physical register to record its status [1]. Typical renaming logic contains Unmap flag, Complete flag, and Counter: Unmap flag indicates whether a register is mapped to an architecture register. Counter is used to record the number of consumers that have not read the information of a register; and Complete flag denotes if a register has been redefined [1].

If Unmap = 1, Complete = 1, and Counter = 0, the register can be released. Accordingly, a short DEAD pulse is generated to turn on N_R and finish discharging process. At the same time, DROWSY is enabled to place the bit-cells to the dead mode. Once a register is mapped (Unmap = 0), DROWSY = DEAD = 0, TM-RF bit-cells enter the work mode.

As discussed in Section III, such conventional register releasing mechanism is designed to support the worst case scenarios and a register may stay in the idle state for many cycles before its Redefiner reclaims. Therefore, a more aggressive implementation is proposed for TM-RF to achieve higher power reduction.

B. Scheme II-High Complexity

This scheme is implemented in conjunction with early register release techniques. As discussed earlier, researchers have explored early register release in two major ways [9], [10]. With hardware support, the first set of solutions release a register when its Redefiner enters the pipeline, instead of waiting for the Redefiner to commit. However, there may be many cycles between a register's LastUser and the dispatch of its Redifiner, missing the opportunity of early releasing. Alternatively, the second set of approaches release registers with compiler support. With compiler analysis, the processor can identify the LastUser and Redefiner of a register, thereby releasing the register earlier.

Therefore, we implement TM-RF based on the combination with compiler assist early register technique [9] [as shown in Fig. 13(b)]. Depending on if there are unresolved branch

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE II
PROCESSOR ORGANIZATION

Reorder buffer	128 entry
Register File	128 integer and 128 floating-point
Machine Width	4-wide fetch, 4-wide issue, 4-wide commit
Load/Store Queue	48 entry load and 48 entry store
Function Units	4 Int ALU, 4 FP ALU, 1 Int MULT/DIV, 1 FP MULT/DIV
BTB	2048 entry, 4-way set-associative
Branch Predictor	Combined with 1024 entries 2-level global predictor with 8 bits history width
L1 I/D Cache	32 KB, 4-way set-associative, 1 cycle hit time
L2 Cache Unified	512 KB, 4-way set-associative, 6 cycles hit time
Memory	64 bit wide, 100 cycles

instructions between a register's LastUser and its Redefiner, two distinct cases arise: 1) if there is no pending branch instruction between the LastUser and Redefiner of a register, once the LastUser reads it, this register can be freed immediately and its bit-cells become the dead mode; and 2) there are still unsolved branch instructions between its LastUser and Redefiner. In this case, the register must hold its data in idle state for recovery from branch misprediction. Since such mispredication happens rarely in practice, its bit-cells can be placed in the drowsy mode and then changes to dead mode when this register can be released safely.

In addition to the conventional register rename logic, another two bits are required to achieve early register release mechanism. Earlyfree flag is set to 1 by the compiler in Case 1 and 0 in Case 2. 1stReady flag is set when the compiler identifies the first instruction which will write valid data in an empty register, thereby putting the register's bit-cells from dead mode back to the work mode. As this instruction commits, 1stReady is cleared. As shown in Fig. 13(b), a DEAD pulse will be generated when a register is freed either in the conventional release condition (Unmap = 1, Complete = 1, and Counter = 0) or in the early release condition (RegUse = 0, Earlyfree = 1, and 1stReady = 0). At the same time, a high DROWSY signal is produced to place bit-cells in dead mode. If Counter = 0 and 1stReady = 0, the register is in idle state and DROWSY = 1 and DEAD = 0, placing the bit-cells in the drowsy mode. Otherwise, DROWSY = DEAD = 0 and the bit-cells become the work mode.

To avoid performance penalty, the TM control logic are operated in parallel with the word-line decoder of RF. Because of their small size, the access delay of TM control logic is much smaller and can be overlapped with the word-line decoding. Therefore, the TM control logic will not create new critical paths in the RF. At the same time, the control logic also introduces a small amount of power overhead. However, the power savings achieved by the proposed technique can offset this overhead, as shown in Section VI-B.

VI. EXPERIMENTAL METHODOLOGY AND RESULTS ON ARCHITECTURAL LEVEL

A. Experimental Methodology

1) Simulator: In addition to the detailed circuit-level implementation of the proposed TM-RF, we carry out architecture-level evaluations by execution-driven simulations using an extensively modified version of the SimpleScalar simulator [34]. In our experiments, we evaluate TM-RF in terms of power efficiency, reliability, temperature profile, area overhead, performance impact, and sensitivity to different RF configurations. Table II describes the processor architecture. We used ten integer SPEC 2000 benchmark suite compiled for the Alpha 21264 processor, based on the reference input set. The benchmarks were simulated after 20 million fast-forward initialization phase.

2) RF Power Model: To evaluate the power savings of the proposed TM-RF, a power model for accessing the RF during a program execution is derived as follows.

Here, we express the power saving of TM-RF over the conventional design as follows:

$$\text{Power}_{\text{saving}} = \left(1 - \sum_i \sum_j \alpha(i) \cdot PS_j(i) \cdot \tau_j(i)\right) \times 100\% \quad (13)$$

where $i = 0, 1$, $j = 1, 2, 3$, $\tau_j(i)$ is the fraction of time that a register containing i is spending in j mode, $\alpha(i)$ is the probability of register containing i, $PS_j(i)$ is the normalized power consumption when register stores i in j mode, and the summation is taken over all possible modes. Note that, to get the overall power consumption of TM-RF in work mode, we consider the number of RF reads is double the number of RF writes [23].

3) Aging Effect: We capture the impact of NBTI/PBTI in the following steps: 1) we obtain the signal probabilities of different devices in RF bit-cells from the extracted stored 0 probability for different applications [see Fig. 5(a)]; 2) to get the temperature profile, we feed the obtained RF power consumption as inputs to the HotSpot thermal simulator [35] with a core floorplan adapted from Alpha 21264 processor; 3) with the obtained signal probability of RF bit-cells and temperature, we use NBTI/PBTI model (1) to calculate the V_{th} shift after the seven-year service time; and 4) we incorporate the calculated V_{th} drifts into 32-nm PTM HK + MG device models and perform circuit-level HSPICE simulations to get performance parameters of RF.

B. Results for Base-Line RFs Configuration

1) Power Efficiency: Based on the RF access statistics obtained from Simplescalar, we estimate the values of $\tau_j(i)$ and $\alpha_j(i)$ in (13) and therefore we can obtain the power saving of the proposed TM-RF.

Fig. 14(a) shows the power reduction achieved by TM-RF compared with the conventional design. The presented results do not consider the power savings of eliminating write 0 operations to empty registers. Here, we also take the dissipations of additional logic of TM-RF into account. For the dissipations in the auxiliary structures, the TM control logic introduces a small power overhead (<0.01%) and thus are not shown; considering two additional bit for each register in the HC scheme, we modeled these bits as traditional SRAM structure and then include their power consumption. As shown

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

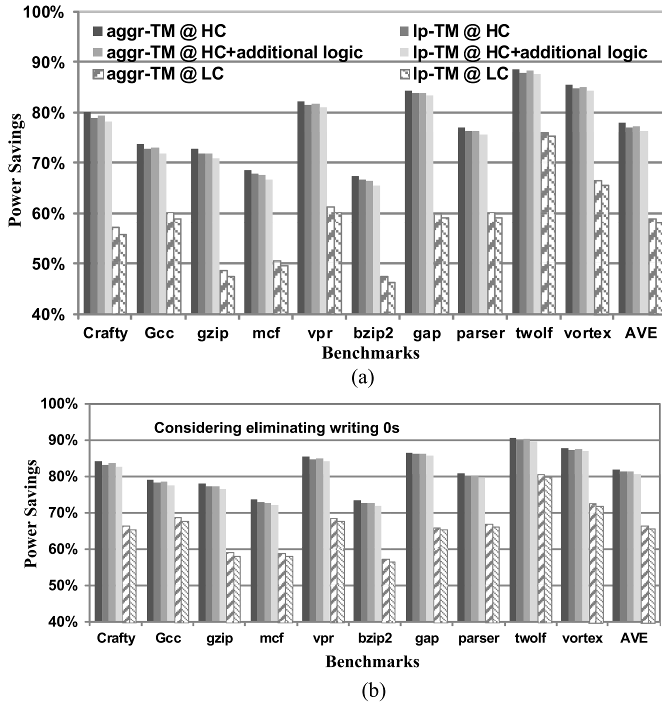


Fig. 14. Power savings of 128-entry 64 b four-read/two-write ported TM-RF over conventional design ($T = 90$ °C).

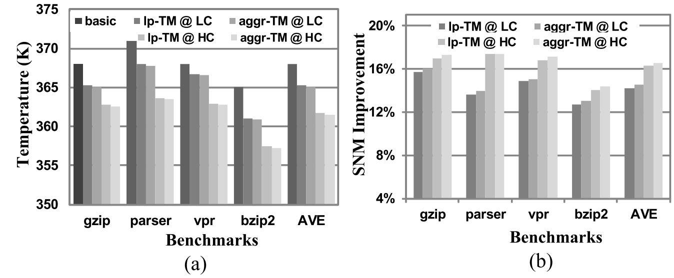


Fig. 15. (a) Temperature reduction. (b) Reliability improvement.

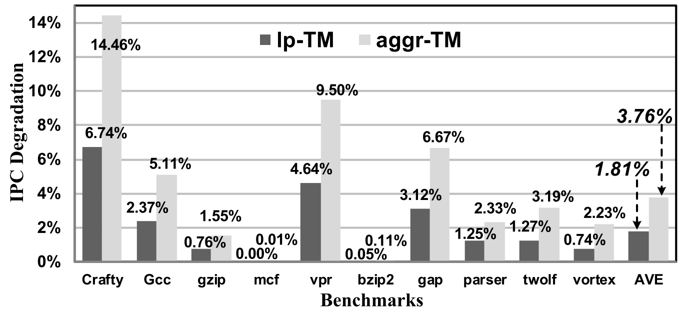


Fig. 16. Performance impact.

in Fig. 14(a), even with the power overhead of additional logic, there are still reductions of 77% (aggr_TM@HC with additional logic), 76% (lp_TM@HC with additional logic), 59% (aggr_TM@LC), and 58% (lp_TM@LC) on average in the RF power consumption. This suggests that by placing RF in different modes, our technique is very effective in reducing RF power consumption.

As discussed in Section IV-A, since a free register contains by default, we can avoid writing a 0 to an empty register with the proposed technique, thereby achieving further power savings of bit-cells, decoder, and word-line drivers. Simulation result in [17] shows that 26% of write operations can be eliminated on average. Accordingly, we reevaluate our proposed technique and the results are shown in Fig. 14(b). As shown, the proposed TM-RF can significantly reduce the power consumed. On average, the power savings of 128-entry 64 b four-read/two-write ported TM-RF can achieve 81.4% (aggr_TM@HC with additional logic), 80.7% (lp_TM@HC with additional logic), 66.3% (aggr_TM@LC), and 65% (lp_TM@LC) compared with basic design.

2) Temperature Profile and Aging Awareness: With Hotspot simulator, we studied the temperature profile of RF while running different applications. Fig. 15(a) shows the steady-state RF temperatures while running four benchmarks (gzip, parser, vpr, bzip2). These benchmarks are four hottest applications and they all increase RF temperature higher than 90 °C. As shown in Fig. 15(a), the temperatures of RF with LC and HC schemes are reduced by 2.9 °C and 6.5 °C on average, respectively. It is worth noting that the reduced temperature can provide extra benefit of bringing down the leakage power, thus further enhancing the power efficiency of TM-RF.

The HS-SNM reliability improvement of TM-RF using different implementation schemes are shown in Fig. 15(b). We can see that, with respect to basic design, lp-TM@LC achieves 14.2%, aggr-TM@LC achieves 14.5%, lp-TM@HC achieves 16.3% reliability improvement after seven service years, and the aggr-TM@HC scheme increases the improvement up to 17%.

3) Area Overhead: Compared with the conventional RF, TM-RF introduces the following new entities each contributing to the overall area overhead: the TM-RF bit-cell and the TM control logic. Compared with the bit-cell overhead, the area of the control logic is much small and can be negligible. As discussed in Section IV-D, the TM-RF bit-cell overhead is 51% for four-read/two-write ported RF. Using CACTI 5.3 tool [24], we obtain that the percentage of bit-cells of the 128-entry 64 b RF area is 16.2%. Thus, the area overhead of the whole register is about 8.2%. In addition, for HC scheme, each register requires additional 2 b and therefore the area overhead is increased to 8.7%.

4) Performance Impact: We evaluate the performance impact of wakeup delay of TM-RF. The transition penalties of lp-TM (one cycle) and aggr-TM (two cycles) are accordingly considered. As shown in Fig. 16, the time overhead is very small [an average instructions per cycle (IPC) loss of 1.81% and 3.76% for lp-TM and aggr-TM]. It is worth mentioning that, in a real microprocessor, typically there is more than a cycle between register renaming of an instruction and its register access, so we expect that the one cycle wakeup delay for lp-TM scheme can be hidden easily in the pipelines, incurring no impact on the performance. Therefore, the proposed TM-RF imposes little performance penalty.

C. Power Reduction Under Different RF Configurations

As discussed in Section IV-C, the configurations of RF also play an important role in determining the effectiveness of

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

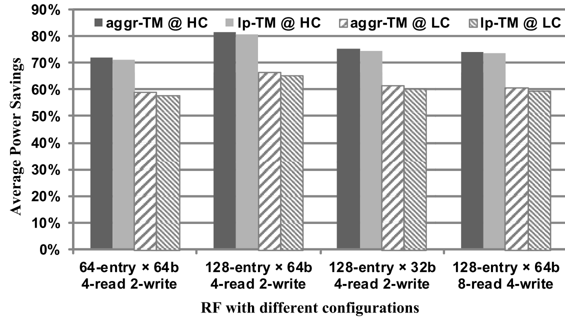


Fig. 17. Average power savings under different RF configurations.

TM-RF. In this section, we study this effect by assessing the power efficiency under different RF configurations in terms of the bits, ports, and size of RF. As shown in Fig. 17, a larger power reduction in RF can be observed with the increase in the number of entries of RF. This is because for a larger RF, the RF occupancy rate decreases, resulting in a larger number of free registers. This enables larger energy reduction according to (13). Meanwhile, as observed in Fig. 17, TM-RF with more bits yields larger power savings. This is because, as number of bits increases, the contribution of bit-line and bit-cell leakage power would be more dominant, enhancing the efficiency of the proposed design. However, when we add more read port and write port to a RF, the peripheral circuits such as word-line decoders, word-line drivers, output drivers and amplifiers constitute a larger portion of RF leakage power consumption. Therefore, the power efficiency of TM-RF, which aims to reduce the power consumption of bit-cells, is decreased, as shown in Fig. 17. Accordingly, to achieve optimized power efficiency in heavily multiported RF, the proposed TM-RF can be applied in conjunction with low-power peripheral techniques, such as the MZZ-HVS approach [32].

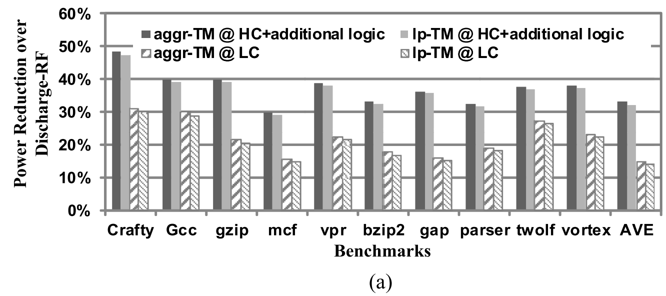
D. Comparison With Existing Low-Power RF Techniques

In this section, we further compare the proposed TM-RF with three existing low-power RF: Discharge-RF [17], CP-RF [16], and Monitor-RF [18].

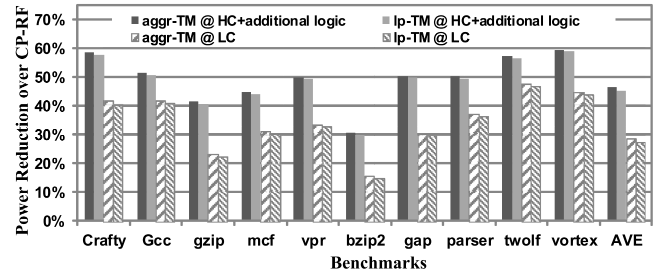
Fig. 18(a) shows the power reduction achieved by TM-RF compared with Discharge-RF for a 128-entry × 64 b RFs. It is shown that the proposed TM-RF is much more effective than Discharge-RF. Specifically, our technique can achieve 15%–33% more power reduction, depending on different implementation schemes. More power saving can be obtained for TM-RF owing to the following reasons.

1) The drowsy mode of TM-RF effectively reduces the power consumption while registers are in idle state.
2) The short-pulse DEAD signal suppresses the gate leakage power, as discussed in Section IV-A.
3) TM-RF bit-cells with high V_{th} write access transistors also lead to lower leakage power.

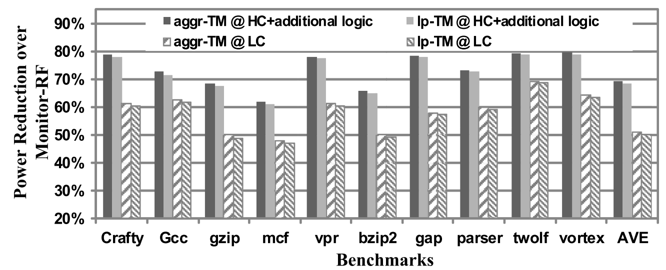
Fig. 18(b) and (c) shows the power reduction of TM-RF over CP-RF and Monitor-RF. We can see that TM-RF significantly outperforms CP-RF and Monitor-RF in average-case power savings by 27%–45% and 50%–69%, respectively, depending on the different implementation schemes.



(a)



(b)



(c)

Fig. 18. Power reduction of the TM-RF compared to prior work (a) Discharge-RF [17], (b) CP-RF [16], and (c) Monitor-RF [18].

We also study the performance of these three designs. To facilitate the comparison, the results of the compared work are all normalized to the case of using conventional RF design. Also, considering the hardware implementation cost, we define a new power efficiency metric, named Efficiency/Cost

$$\text{Efficiency/Cost} = \frac{\text{Power_savings} \times \text{IPC}}{\text{Area_overhead}}. \tag{14}$$

Table III shows the comparison among different RF designs. We can see that the Efficiency/Cost of CP-RF, Monitor-RF, and Discharge-CF are 0.31, 0.06, and 0.415, respectively. In comparison, our design can achieve a high Efficiency/Cost of 0.52–0.69. In addition, the proposed TM-RF successfully improves the reliability by 14.2%–17%, while with only 1.81%–3.76% performance degradation on average, respectively. The results clearly show that the proposed TM-RF consistently outperforms the existing solutions in terms of both power efficiency and reliability. Moreover, the proposed TM-RF design with four schemes is more flexible to handle different types of applications. Therefore, the proposed TM-RF can be a preferable solution for implementing power-efficient and reliable RF in modern microprocessors.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE III
COMPARISON WITH EXISTING LOW-POWER RF

	Discharge-RF [17]	CP-RF [16]	Monitor-RF [18]	**This work**			
				lp-TM@LC	**aggr-TM@LC**	**lp-TM@HC**	**aggr-TM@HC**
Bit-cell	Modified	Modified	Basic	Modified			
Bit-cell Area	~1.30x	~1.62x[1]	1x	1.51x			
Additional hardware	No	2 bits per physical register and 2 bits per rename entry	DVS controller	TM-controller		TM controller and 2 bits per physical register	
Total Area Overhead	5.9% RF	~11% RF	20% (RF+ROB)	~8.2% RF		~8.7% RF	
Performance (IPC)	~1	~1.109x	~90%	98.19%	96.24%	98.19%	96.24%
Aging awareness (improvement)	-	-	-	Yes (14.2%)	Yes (14.5%)	Yes (16.3%)	Yes (17%)
Efficiency/Cost	0.415	0.31	0.06	0.526	0.522	0.69	0.682

[1] Layout overhead is scaled to 128-entry × 64b 4-read/2-write ported RF.

VII. CONCLUSION

In this paper,we have presented a trimodal RF technique for modern microprocessors. The technique employs a trimodal bit-cell and relates these three modes to states of registers, thereby reducing RF power consumption. We developed four schemes to implement TM-RF, providing flexibility for different applications. We used device selection (high V_{th} write access devices) and worst case sizing methodology to mitigate aging-effect-induced reliability degradation. Simulation results demonstrate significant reduction in power consumption and noticeable reliability mitigation with minimal area overhead, while maintaining almost the same performance as compared with the conventional design. Our future investigations would include extension of the proposed TM-RF technique to deal with multithreaded workloads.

REFERENCES

[1] M. Moudgill, K. Pingali, and S. Vassiliadis, "Register renaming and dynamic speculation: An alternative approach," in *Proc. 26th Annu. Int. Symp. Microarchit.*, Dec. 1993, pp. 202–213.

[2] G. S. Ditlow *et al.*, "A 4R2W register file for a 2.3 GHz wire-speed POWER processor with double-pumped write operation," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Paper (ISSCC)*, Feb. 2011, pp. 256–258.

[3] J. L. Shin *et al.*, "The next-generation 64b SPARC core in a T4 SoC processor," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Paper (ISSCC)*, Feb. 2012, pp. 60–62.

[4] N. Gong, J. Wang, S. Jiang, and R. Sridhar, "Clock-biased local bit line for high performance register files," *Electron. Lett.*, vol. 48, no. 18, pp. 1104–1105, Aug. 2012.

[5] X. Guan and Y. Fei, "Register file partitioning and compiler support for reducing embedded processor power consumption," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 18, no. 8, pp. 1248–1252, Aug. 2010.

[6] L. Li, Y. Zhang, and J. Yang, "Proactive recovery for BTI in high-*k* SRAM cells," in *Proc. Design, Autom. Test Eur. Conf. Exhibit. (DATE)*, Mar. 2011, pp. 992–997.

[7] J. A. Blome, S. Gupta, S. Feng, and S. Mahlke, "Cost-efficient soft error protection for embedded microprocessors," in *Proc. Int. Conf. Compil., Archit. Synth. Embedded Syst.*, 2006, pp. 421–431.

[8] N. Gong, G. Tang, J. Wang, and R. Sridhar, "Low power tri-state register files design for modern out-of-order processors," in *Proc. IEEE Int. Syst. Chip Conf. (SoCC)*, Sep. 2011, pp. 323–328.

[9] T. M. Jones, M. F. P. O'Boyle, J. Abella, A. González, and O. Ergin, "Exploring the limits of early register release: Exploiting compiler analysis," *ACM Trans. Archit. Code Optim.*, vol. 6, no. 3, pp. 12–30, Sep. 2009.

[10] T. Monreal, V. Vinals, A. Gonzalez, and M. Valero, "Hardware schemes for early register release," in *Proc. Int. Conf. Parallel Process. (ICPP)*, 2002, pp. 5–13.

[11] T. Monreal, V. Vinals, J. Gonzalez, A. Gonzalez, and M. Valero, "Late allocation and early release of physical registers," *IEEE Trans. Comput.*, vol. 53, no. 10, pp. 1244–1259, Oct. 2004.

[12] R. Sangireddy and A. K. Somani, "Exploiting quiescent states in register lifetime," in *Proc. IEEE Int. Conf. Comput. Design, Very Large Scale Integr. (VLSI) Comput. Process. (ICCD)*, Oct. 2004, pp. 368–374.

[13] S. Part, A. Shrivastava, N. Dutt, A. Nicolau, Y. Paek, and E. Earlie, "Bypass aware instruction scheduling for register file power reduction," in *Proc. ACM SIGPLAN/SIGBED Conf. Lang., Compil., Tool Support Embedded Syst. (LCTES)*, 2006, pp. 173–181.

[14] D. Balkan, J. Sharkey, D. Ponomarev, and K. Ghose, "Selective writeback: Reducing register file pressure and energy consumption," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 16, no. 6, pp. 650–661, Jun. 2008.

[15] R. Shioya, K. Horio, M. Goshima, and S. Sakai, "Register cache system not for latency reduction purpose," in *Proc. 43rd Annu. IEEE/ACM Int. Symp. Microarchit. (MICRO)*, Dec. 2010, pp. 301–312.

[16] O. Ergin, D. Balkan, D. Ponomarev, and K. Ghose, "Early register deallocation mechanisms using checkpointed register files," *IEEE Trans. Comput.*, vol. 55, no. 9, pp. 1153–1164, Sep. 2006.

[17] L. Jin, W. Wu, J. Yang, C. Zhang, and Y. Zhang "Reduce register files leakage through discharging cells," in *Proc. Int. Conf. Comput. Design (ICCD)*, Oct. 2007, pp. 114–119.

[18] W.-Y. Shieh and H.-D. Chen, "Saving register-file static power by monitoring instruction sequence in ROB," *J. Syst. Archit.*, vol. 57, no. 4, pp. 327–339, Apr. 2011.

[19] S. Kothawade, K. Chakraborty, and S. Roy, "Analysis and mitigation of NBTI aging in register file: An end-to-end approach," in *Proc. 12th Int. Symp. Qual. Electron. Design (ISQED)*, Mar. 2011, pp. 1–7.

[20] H. Mostafa, M. Anis, and M. Elmasry, "Adaptive body bias for reducing the impacts of NBTI and process variations on 6T SRAM cells," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 58, no. 12, pp. 2859–2871, Dec. 2011.

[21] T. Siddiqua and S. Gurumurthi, "Enhancing NBTI recovery in SRAM arrays through recovery boosting," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 20, no. 4, pp. 616–629, Apr. 2012.

[22] N. Gong, S. Jiang, J. Wang, B. Aravamudhan, K. Sekar, and R. Sridhar, "Hybrid-cell register files design for improving NBTI reliability," *Microelectron. Rel.*, vol. 52, nos. 9–10, pp. 1865–1869, Sep./Oct. 2012.

[23] S. Roy, N. Ranganathan, and S. Katkoori, "State-retentive power gating of register files in multicore processors featuring multithreaded in-order cores," *IEEE Trans. Comput.*, vol. 60, no. 11, pp. 1547–1560, Nov. 2011.

[24] (2008). *CACTI5*, Hewlett-Packard Company, Palo Alto, CA, USA. [Online]. Available: http://quid.hpl.hp.com:9081/cacti

[25] H. Homayoun, A. Sasan, J.-L. Gaudiot, and A. Veidenbaum, "Reducing power in all major CAM and SRAM-based processor units via centralized, dynamic resource size management," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 19, no. 11, pp. 2081–2094, Nov. 2011.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[26] (2008). *Predictive Technology Model (PTM)*. [Online]. Available: http://www.eas.asu.edu/~ptm
[27] N. S. Kim, S. C. Draper, S.-T. Zhou, S. Katariya, H. R. Ghasemi, and T. Park "Analyzing the impact of joint optimization of cell size, redundancy, and ECC on low-voltage SRAM array total area," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 20, no. 12, pp. 2333–2337, Dec. 2012.
[28] A. Agarwal, H. Li, and K. Roy, "A single-V_t low-leakage gated-ground cache for deep submicron," *IEEE J. Solid-State Circuits*, vol. 38, no. 2, pp. 319–328, Feb. 2003.
[29] H.-I. Yang, S.-C. Yang, W. Hwang, and C.-T. Chuang "Impacts of NBTI/PBTI on timing control circuits and degradation tolerant design in nanoscale CMOS SRAM," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 58, no. 6, pp. 1239–1251, Jan. 2011.
[30] S. Zafar *et al.*, "A comparative study of NBTI and PBTI (charge trapping) in SiO_2/HfO_2 stacks with FUSI, TiN, Re gates," in *Symp. Very Large Scale Integr. (VLSI) Technol. Dig. Tech. Papers*, Oct. 2006, pp. 23–25.
[31] A. Tiwari and J. Torrellas, "Facelift: Hiding and slowing down aging in multicores," in *Proc. 41st IEEE/ACM Int. Symp. Microarchit.*, Nov. 2008, pp. 129–140.
[32] H. Homayoun, A. Sasan, A. Veidenbaum, H.-C. Yao, S. Golshan, and P. Heydari, "MZZ-HVS: Multiple sleep modes zig-zag horizontal and vertical sleep transistor sharing to reduce leakage power in on-chip SRAM peripheral circuits," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 19, no. 12, pp. 2303–2316, May 2011.
[33] (2009). *MOSIS Deep Design Rules*. [Online]. Available: http://www.mosis.com/
[34] D. Burger and T. M. Austin, "The SimpleScalar tool set: Version 2.0," Dept. Comput. Sci., Univ. Wisconsin-Madison, Madison, WI, USA, Tech. Rep. CS-1342, 1997.
[35] W. Huang, S. Ghosh, S. Velusamy, K. Sankaranarayanan, K. Skadron, and M. R. Stan, "HotSpot: A compact thermal modeling methodology for early-stage VLSI design," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 14, no. 5, pp. 501–524, May 2006.
[36] N. Gong, B. Guo, J. Lou, and J. Wang, "Analysis and optimization of leakage current characteristics in sub-65 nm dual V_t footed domino circuits," *Microelectron. J.*, vol. 39, no. 9, pp. 1149–1155, Sep. 2008.
[37] (2010). *International Technology Roadmap for Semiconductors (ITRS)*. ITRS, London, U.K. [Online]. Available: http://www.itrs.net
[38] R. Vattikonda, W. Wang, and Y. Cao, "Modeling and minimization of pMOS NBTI effect for robust nanometer design," in *Proc. 43rd ACM/IEEE Design Autom. Conf.*, 2006, pp. 1047–1052.
[39] S. V. Kumar, C. H. Kim, and S. S. Sapatnekar, "An analytical model for negative bias temperature instability (NBTI)," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design*, Nov. 2006, pp. 493–496.
[40] J. Chen, S. Wang, and M. Tehranipoor, "Efficient selection and analysis of critical-reliability paths and gates," in *Proc. Great Lakes Symp. Very Large Scale Integr. (VLSI)*, 2012, pp. 45–50.
[41] W. Wang, V. Balakrishnan, B. Yang, and Y. Cao, "Statistical prediction of NBTI-induced circuit aging," in *Proc. Int. Conf. Solid-State Integr.-Circuit Technol.*, Oct. 2008, pp. 416–419.
[42] H.-I. Yang, W. Hwang, and C.-T. Chuang, "Impacts of NBTI/PBTI and contact resistance on power-gated SRAM with high-*k* metal-gate devices," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 19, no. 7, pp. 1192–1204, Jul. 2011.

Na Gong (M'13) received the B.E. degree in electrical engineering and the M.E. degree in microelectronics from Hebei University, Hebei, China, in 2004 and 2007, respectively, and the Ph.D. degree in computer science and engineering from the State University of New York, Buffalo, NY, USA, in 2013.

She is currently an Assistant Professor of Electrical and Computer Engineering with North Dakota State University, Fargo, ND, USA. Her current research interests include device-circuit-architecture codesign for nanoscale very large scale integration circuit and system, power efficient and reliable electronics for mobile computing and high-performance computing, and emerging memory technologies in computer systems.

Jinhui Wang (M'13) received the B.E. degree in electrical engineering from Hebei University, Hebei, China, in 2004, and the Ph.D. degree in electrical engineering through a joint USA/China program between University of Rochester and Beijing University of Technology, in 2010.

Dr. Wang is currently an Associate Professor with the College of Electronic Information and Control Engineering, Beijing University of Technology, Beijing, China. His research interests include low-power, high-performance, and variation-tolerant integrated circuit design, 3-D IC and EDA methodologies, and thermal issue solution in VLSI. He has more than 70 publications and 6 patents in the emerging semiconductor technologies.

Shixiong Jiang received the M.S. degree in electrical engineering from the University at Buffalo (UB), Buffalo, NY, USA, in 2012.

He joined the Very Large Scale Integration (VLSI) Laboratory at UB in 2012. His current research interests include VLSI, computer architecture, and EDA.

Ramalingam Sridhar (M'82–SM'99) received the B.E. (honors) degree in electrical and electronics engineering from Guindy Engineering College, University of Madras, Chennai, India, and the M.S. and Ph.D. degrees in electrical and computer engineering from Washington State University, Pullman, WA, USA.

He has been with the State University of New York, Buffalo, NY, USA, since 1987, where he is an Associate Professor with the Department of Computer Science and Engineering. His current research interests include very large scale integration (VLSI) circuits, systems and architecture, variability, power-aware and robust design, very deep submicrometer systems, deep submicrometer VLSI systems, clocking and synchronization, memory circuits and architecture, wireless and sensor network security, secure architectures, and power-aware security solutions for embedded systems.

Prof. Sridhar was an IEEE CAS Distinguished Lecturer. He has served as the Program Chair and General Chair of the Application Specific Integrated Circuits and the International System-on-Chip Conferences, and has served on the Editorial Board of many journals, including the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—PART I: REGULAR PAPERS and the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION SYSTEMS, and the technical committees of numerous conferences in wireless systems and VLSI.