Register file reliability enhancement through adjacent narrow-width exploitation.

Hamzeh Ahangari, Ihsen Alouani, Özcan Özturk, Smail Niar, Atika Rivenq

Register File Reliability Enhancement Through Adjacent Narrow-width Exploitation

Hamzeh Ahangari[1], Ihsen Alouani[2], Ozcan Ozturk[1], Smail Niar[2], and Atika Rivenq[3]

[1]Department of Computer Engineering, Bilkent University, Ankara, Turkey
[2]LAMIH lab, University of Valenciennes, France
[3]IEMN/DOAE lab, University of Valenciennes, France

Abstract—Due to the increasing vulnerability of CMOS circuits, new generations of microprocessors require an inevitable focus on reliability issues. As the Register File (RF) constitutes a critical element within the processor pipeline, it is mandatory to enhance the RF reliability to develop fault tolerant architectures. This paper proposes Adjacent Register Hardened RF (ARH), a new RF architecture that exploits the adjacent byte-level narrow-width values for hardening registers at runtime. Registers are paired together by some special switches referred to as joiners. Dummy sign bits of each register are used to keep redundant data of its counterpart register. We use 7T/14T SRAM cell [6] to combine redundant bits together to make a single bit cell which is, by far, more resilient against faults. Our simulations show that with 3% to 12% power overhead and 10% to 20% increase in area, in comparison to baseline RF, we can obtain up to 80% reduction in soft error rate (SER).

Keywords: Register file, reliability, soft error rate, SER, narrow-width, 7T/14T SRAM.

I. INTRODUCTION

In recent years, as sub-micron technology dimensions sharply decreased to a few nanometer range, new types of challenges are introduced. Reliability of electronic circuits is one such concern which calls for more investigation. Since microprocessors are becoming more vulnerable to various types of faults than past.

By growing chip density, Soft Error Rates (SER) per system grows. Similarly in newer technologies, particles with lower energy are able to induce fault causing an increase in SER. Moreover, protecting microprocessors memory and sequential elements is critical because of its direct impact on systems reliability and data correctness. Cache memory, register file (RF), flipflop (FF) and latch are usual sequential parts of a microprocessor architecture, each of which requires its own suitable solutions for reliability enhancement. Both cache and RF are based on SRAM memory structure. However, since their characteristics and applications differ, their prevalent reliability techniques also differ. In caches, ECC is an effective technique for protection against faults. However, unlike cache, due to timing and power overheads, ECC is not an appropriate solution for register file reliability. In RF, activity rate per address is higher than cache memories, making power consumption more important. Additionally, RF is in processor's critical path and priority of performance is an essential necessity. Consequently, finding suitable technique for RF reliability enhancement is a new kind of challenge when compared to cache memories.

In this paper, we propose a relatively different approach to exploit vacant spaces in RF to keep redundant data. This novel idea combines both architectural and circuit techniques to achieve more robustness in RF. Provided that one of two SRAM cells is vacant, meaning it is filled with a dummy sign bit, those two SRAM cells are joined together in circuit level by means of two transistors to make one more robust SRAM cell. The signals to apply such joining is issued by reliability control unit.

II. RELATED WORK

In some studies, register duplication is proposed. For example, in [11] by means of register renaming unit, unused registers are detected and exploited to preserve redundant copies of other registers.

In-Register Duplication (IRD) is proposed in [8], [9] in which, by an opportunistic idea, dummy sign bits of narrow-width register values are replaced with replication of meaningful bits during RF write operation. In read operation, replicated and original bits are bitwise compared to find mismatch as error indication. Additionally, two parity bits are embedded for each half. By means of both error detection mechanisms, which together are similar to a 2D parity system, they added error detection/recovery for narrow width values stored in RF. Nevertheless, long operands are not protected by IRD. If applied to 32-bit RF, this disadvantage is more serious, because long operands are frequent.

All of the above-mentioned works are architectural level ideas based on information redundancy and explicit comparison operation. The main difference of our work is that we combine a circuit level hardening technique with narrow width duplication. In addition to reducing SER, by clever replication in two paired registers, unlike IRD works [8], [9], we protect long operands better than previous works. Provided that a long operand is next to a short one, priority is given to long operand and replication of its more significant bits are done on dummy sign bits of short operand.
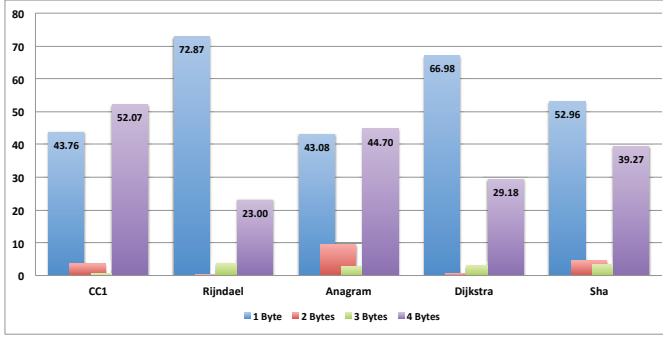
Fig. 1. Distribution of effective length of numbers in 32-bit RF in some benchmarks.
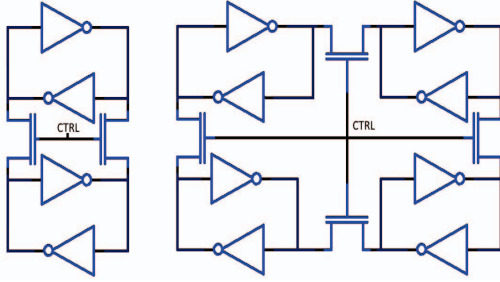

Fig. 2. Left: 7T/14T memory cell with nMOS joiners [6] right: JSRAM cell with nMOS joiners [1].

III. OUR ARCHITECTURE

Our approach tries to improve reliability of RF by exploiting unused bits of integer numbers in adjacent registers for hardening cells. For any number in range of minimum to maximum possible values in 2's complement system, only one single sign bit is sufficient for correct representation of the number. The remaining sign bits are just multiple copies of the same sign bit and are vain redundant bits. Based on this, instead of preserving multiple redundant bits for sign, we suggest to exploit them to enhance the reliability of adjacent registers. Adjacent Register Hardening (ARH) is very efficient to protect highly critical data within an application using dummy bits of non-critical registers. Since the content of registers are unveiled at run time, the extent of reliability increase is application dependent. Figure 1 shows that, on average, numbers with effective length of one byte constitute more than half of the numbers stored in the RF, in the tested benchmarks.

The implementation consists of retrieving the data to be stored, the technical solution to enhance reliability and perform the different read/write access. As detailed in next section, instead of relying merely on high-level architectural solutions, in our implementation we get benefit from a fast circuit level technique combined with higher-level architectural control, to build a highly flexible reliability solution.

A. Circuit Level Reliability Enhancement

7T/14T [6] proposed combining two SRAM cells in circuit level to achieve more reliability or performance dynamically (Figure 2 left). According to this idea, two memory cells are joined upon request to store single bit of data. Joining is done by activating two transistors that connect the internal nodes of two cells to each other. For biasing toward reliability and not performance, just one of the two wordline signals is used for read or write operation [6]. If joiners are not activated (CTRL="L", if switches are nMOS), then the proposed structure works normally as two separated conventional 6T SRAM cells. JSRAM cell [1] is an extension of 7T/14T cell to combine four cells in a ring fashion to achieve full immunity against single bit errors by providing an auto correction mechanism (Figure 2 right). It is also capable of tolerating multiple bit upsets (MBUs). Since the reliability enhancement in our current work is in a statistical way and is dependent on the values stored in registers, using 7T/14T cell is more justified.

In our proposed architecture, adjacent registers of RF are joined together by 7T/14T technique. Generally, each bit can be joined to any number of bits from any register, by embedding multiple switches in between. Nevertheless, to avoid excessive area overhead and complexity, we limit this idea by just allowing each bit to be joined into a unique bit of a specific register. Thus, registers are paired together, bit by bit, during RF design. The benefit of pairing non-adjacent registers would be less probability of being affected by MBU, but obviously with more routing overhead. Currently we opt to combine neighbor registers.

B. Architecture Level Organization

To enhance the RF error resiliency, we take advantage of the reconfigurable aspect of the 7T/14T cell. Instead of relying on ECCs or extra memory space for reliability enhancement, we opt for an opportunistic approach that exploits unused bits within the stored data. To optimize the reconfiguration circuitry, as well as the additional bit cells, we opted for a byte-level granularity. Accordingly, the idle bytes are used to harden registers against errors.

Considering byte level granularity, a judicious one-to-one mapping between bytes of two registers is required to exploit the empty bits efficiently. Dummy sign bits are on the left-hand side (MSB side), while real data bits are on the other side. Thus, first obvious paradigm of mapping is in a crossed way, byte-0 of one register to byte-3 of the paired register, byte-2 to byte-1 and so on. However we've taken into account a second point in byte mapping. Faults in more valuable bits of an integer, lead to more absolute numerical error. While the best mapping is application dependent, we extracted the distribution of operand length for our benchmarks as had been shown in Figure 1. Operands with lengths of one byte and four bytes are dominant ones. Then paired registers of length one-one, one-four and four-four are more frequent. This means byte mapping has to be biased toward protecting one-one and one-four combinations (four-four can not be protected). Hence,
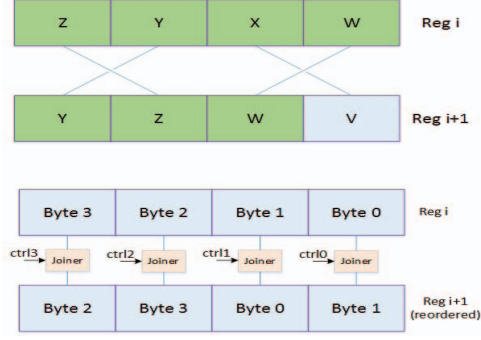
Fig. 3. Top: Three bytes of "ZYXW" number in reg-i are replicated in sign bits of reg-i+1. "V" number in reg-i+1 is not replicated. Bottom: easy routing by byte reordering.

by limiting ourselves to at most four groups of byte-to-byte joiners, we take mapping of Figure 3 as most efficient one which leads to better RF error resiliency. For a 32-bit RF, four control signals are required for controlling this mapping.

One superiority of our work in comparison to In-Register Duplication (IRD) works is that, by pairing registers, ARH can protect long operands. For example, in Figure 3, reg-i occupies four bytes and three of these bytes are protected by reg-i+1 which occupies only one byte. However in IRD, long operands which represent larger integers are not protected.

Below, we describe the mechanism for basic write/read operations:

1) Write Access: Mechanism behind the write operation is critical to achieve efficiency. During write operation, only meaningful bytes are written, while dummy sign bits should not be written and respective bytes in register are left intact. Because those bytes may be keeping the redundant data of the other paired register. This can be satisfied by having byte selectable write enables. Besides this, when those meaningful bytes are being written, while their counterpart bytes in the other register are not in use, in this situation control signal of joiners have to be activated. According to electrical characteristics of 7T/14T cell, if joiner is activated and one of the paired cells is written, the other one is written automatically as well. By exploiting this property, by single write operation, redundant data is quickly written at the same time into the redundant byte of the other paired register.

Above-mentioned mechanism requires modification to ALU and RF decoder. The ALU should simply detect effective length of integer numbers. In addition to storing data within the targeted register address, 2-bit effective length value (EL) is also stored beside the register (Figure 4). Considering EL value, only write enable signals of necessary bytes are activated, allowing writing the data with size of effective length into register.

By means of available EL value of the paired register (paired register of the register which is being written), unused bytes of paired register are determined to store redundant data. Then proper control signals are generated by a simple two-level

AND-OR circuit.

During the write access, the reliability controller unit sets the configuration to adapt the available idle bytes to protect the data which is being written. Although extra circuitry of reliability controller is on critical path, by combining it with decoder during logic synthesis, the delay overhead is minimized. For easier routing, bytes of one of registers can be reordered. The required multiplexer is in parallel with decoder and not inside critical path (Figure 4 left).

2) Read Access: The read access architecture is modified to cope with the reliability enhancement process. Once a register's idle bytes are exploited for hardening cells, they should be replaced with actual sign value during the read access to insure data integrity.

As shown in Figure 4 right, the reliability controller unit selects whether the forwarded data would be the "directly read byte" or the "sign byte", depending on the register effective width. If the byte-reordering has been already employed in write operation, actual order have to be recovered again. To avoid the timing overhead of sign bit detection, sign bit can also be stored explicitly like EL in write operation. Otherwise, sign bit has to be determined by finding MSB bit of most significant byte in read operation. All these are performed by a multiplexer as depicted in Figure 4 right. This multiplexer selects one of four inputs: directly read byte, reordered byte, all 0/1 for sign extension of positive/negative numbers.

IV. EXPERIMENTS

To confirm the circuit functionality and calculate area and power overheads, simulation with HSPICE was performed with 22nm predictive technology model library [12]. Transistor sizes for typical 22nm SRAM cell were chosen from [13]. Ratio values are: cell ratio $= WPD/WPG = 2.02$ and pullup ratio $= WPU/WPG = 1.18$. Wordline pulse width is chosen as 1ns.

We selected typical values of original SER and improved rates using 7T/14T SRAM cell form [5] and [15]. Although those experimental results are related to SRAM chips fabricated in different technologies (65nm and 150nm), we only considered the improvement ratios, not the exact values, as an approximation. Although SER per system increases sharply by technology size reduction, but SER per memory bit grows gently [3]. Therefore, sensitivity of improvement to technology is not expected to be high.

In this section, the system-level experimentations are presented for a typical 32 x 32 bit register file, where power oriented experiments were conducted. In order to get accurate simulation results, a WATTCH power simulator [4] was modified by estimating the cycle-accurate power consumption using HSPICE results. Hence, cycle-level simulations based on a 5-stage pipeline out-of-order processor modeled by a SimpleScalar simulation environment [2] were performed. We extensively modified the simulator code to support the proposed reliability enhancement technique.

For this evaluation, benchmarks from two different sets of applications, namely the SPEC CPU2000 benchmark suite [14]
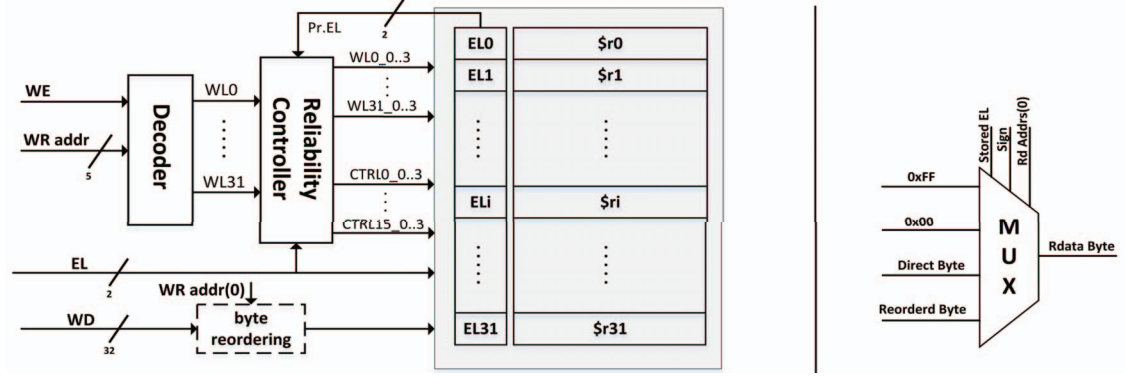
Fig. 4. Left: Write Access Circuit, Wordline and Joiner Signals Right: Read Access Multiplexer
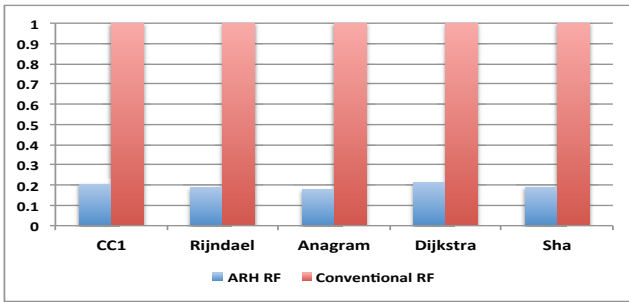

Fig. 5. Normalized error rate of ARH RF vs conventional RF.

and MiBench [7], were compiled for the Alpha instruction set architecture.

To evaluate the error resilience of ARH RF, we developed an exhaustive fault injection platform where the injected error locality is randomly defined. Considering again the benchmark distribution depicted in Figure 1, and referred SER of protected and unprotected bits, normalized error rates are shown in Figure 5.

The increase in static power because of joiner switches is negligible. Our power consumption simulations show that the overall power overhead does not exceed 12% in the worst case.

Operand detection circuit is very simple and it has no effect on latency. For each pair of bits, two switches are added in between them. Depending on type of switches and number of read and write ports of RF, area overhead is indicated to be around 10%-20% [5].

V. Conclusion and future work

In this work we proposed a novel narrow-width register duplication technique. By a new approach, we exploit dummy sign bits for hardening data bit cells at circuit level, benefiting from configurable 7T/14T SRAM cell structure. According to the proposed technique, adjacent registers are paired together. Nonsignificant bits of one register are exploited for reliability enhancement of the other register. This aspect not only affords protection to long-length values but also is very efficient in critical data protection. Results show that by benefiting from considerable SER improvement of 7T/14T and a judicious byte pairing, error rates of integers stored in RF are reduced significantly in comparison to baseline RF.

References

[1] Ahangari H, Yalcin G, Ozturk O, Unsal O, Cristal A. JSRAM: A Circuit-Level Technique for Trading-Off Robustness and Capacity in Cache Memories. InVLSI (ISVLSI), 2015 IEEE Computer Society Annual Symposium on 2015 Jul 8 (pp. 149-154). IEEE.

[2] Austin T, Larson E, Ernst D. SimpleScalar: An infrastructure for computer system modeling. Computer. 2002 Feb;35(2):59-67.

[3] Baumann R. Soft errors in advanced computer systems. Design and Test of Computers, IEEE. 2005 May;22(3):258-66.

[4] Brooks D, Tiwari V, Martonosi M. Wattch: a framework for architectural-level power analysis and optimizations. ACM; 2000 Jun 10.

[5] Fujiwara H, Okumura S, Iguchi Y, Noguchi H, Kawaguchi H, Yoshimoto M. A dependable SRAM with 7T/14T memory cells. IEICE transactions on electronics. 2009 Apr 1;92(4):423-32.

[6] Fujiwara H, Okumura S, Iguchi Y, Noguchi H, Morita Y, Kawaguchi H, Yoshimoto M. Quality of a bit (QoB): A new concept in dependable SRAM. InQuality Electronic Design, 2008. ISQED 2008. 9th International Symposium on 2008 Mar 17 (pp. 98-102). IEEE.

[7] Guthaus MR, Ringenberg JS, Ernst D, Austin TM, Mudge T, Brown RB. MiBench: A free, commercially representative embedded benchmark suite. InWorkload Characterization, 2001. WWC-4. 2001 IEEE International Workshop on 2001 Dec 2 (pp. 3-14). IEEE.

[8] Hu J, Wang S, Ziavras SG. In-register duplication: Exploiting narrow-width value for improving register file reliability. InDependable Systems and Networks, 2006. DSN 2006. International Conference on 2006 Jun 25 (pp. 281-290). IEEE.

[9] Hu J, Wang S, Ziavras SG. On the exploitation of narrow-width values for improving register file reliability. Very Large Scale Integration (VLSI) Systems, IEEE Transactions on. 2009 Jul;17(7):953-63.

[10] Kandala M, Zhang W, Yang LT. An area-efficient approach to improving register file reliability against transient errors. InAdvanced Information Networking and Applications Workshops, 2007, AINAW'07. 21st International Conference on 2007 May 21 (Vol. 1, pp. 798-803). IEEE.

[11] Memik G, Kandemir MT, Ozturk O. Increasing register file immunity to transient errors. InDesign, Automation and Test in Europe, 2005. Proceedings 2005 Mar 7 (pp. 586-591). IEEE.

[12] Predictive technology model, http://ptm.asu.edu

[13] Shin C. Advanced MOSFET designs and implications for SRAM scaling (Doctoral dissertation, University of California, Berkeley).

[14] Spec cpu2000 benchmarks, http://www.spec.org/cpu2000/index.html

[15] Yoshimoto S, Amashita T, Okumura S, Yamaguchi K, Yoshimoto M, Kawaguchi H. Bit error and soft error hardenable 7T/14T SRAM with 150-nm FD-SOI process. InReliability Physics Symposium (IRPS), 2011 IEEE International 2011 Apr 10 (pp. SE-3). IEEE.