|  |  |
| --- | --- |
| In-Person: Oral [x] / Poster [ ] / The same [ ] Virtual: Pre-recorded video or Zoom [ ] *<please select the type of your presentation>* | Topic: *<Microelectronics reliability and qualification>* |

Effect of Asymmetric Transistor Aging on GPGPUs

**F. Gabbay1, F. Ramadan2, M. Ganaiem2, Ofrie Rosenthal2 and Lior Bashari2**
1 Engineering Faculty, Ruppin Academic Center, 4025000, Emek Hefer, Israel

2 Electrical and Computer Engineering Faculty, Technion – Israel Institute of Technology, Technion City, 3200000 Haifa, Israel
E-mail: {freddyg@ruppin.ac.il, firasramadan@campus.technion.ac.il, majd.ga@campus.technion.ac.il, ofrie.r@campus.technion.ac.il, li-orbashari@campus.technion.ac.il}

**Summary:** General-purpose graphics processing units (GPGPUs) are specialized hardware devices designed for parallel computing tasks in various domains. Their integration into critical systems such as autonomous vehicles, security systems, and medical devices has increased demand for enhanced reliability and resilience to satisfy industry and regulatory standards. However, GPGPUs face reliability concerns due to transistor aging, which is caused by bias-temperature instability (BTI). This progressive degradation of transistor performance can degrade performance and cause critical circuit failures, producing timing violations. The present work investigates how transistor aging affects GPGPU execution units, highlighting their vulnerability to BTI. Experimental analysis reveals that BTI significantly influences computational elements within GPGPUs. To address this issue, we propose a mitigation technique that specifically targets the challenges of asymmetric aging in GPGPU execution units, thereby mitigating timing violations.

**Keywords:** GPGPU, BTI, Asymmetric transistor aging

**1. Introduction**

General-purpose graphics processing units (GPGPUs) are specialized hardware devices designed to perform highly parallel computations [1,2] to accelerate various computational tasks, including machine learning, high-performance computing, scientific simulations, data analytics, and more. Furthermore, GPGPUs have recently seen heavy use in critical systems such as autonomous vehicles, security systems, and medical devices [3,4]. These emerging applications impose stringent requirements on the resiliency and reliability of GPGPUs, as mandated by industry and regulatory standards.

In recent decades, VLSI technologies have profited from remarkable advancements following several significant trends. First, the continuous development of new process nodes has ensured the consistent miniaturization of transistors to nanometric dimensions, in line with Moore’s law. Second, revolutionary devices and materials have been pivotal in driving advancements, resulting in improved performance and reduced energy consumption. However, these advancements have also highlighted the vulnerability of integrated circuits (ICs) to reliability issues, particularly those caused by transistor aging [5,6]. Transistor aging refers to the decline in a transistor’s performance over time, primarily due to the bias temperature instability (BTI), as further described in Section 2. The BTI significantly affects IC reliability, causing performance degradation and critical circuit failures due to timing violations. Moreover, asymmetric aging exacerbates timing violations and amplifies reliability concerns because it results from unevenly distributed degradation.

The present work investigates how transistor aging affects GPGPU execution units. Our experimental analysis includes functional and physical simulations, indicating that computational elements within GPGPUs can be highly susceptible to BTI. Additionally, our analysis indicates that the various execution units within GPGPU processing elements (PEs) may experience asymmetric aging, resulting in even more serious timing violations. Thus, we present herein a mitigation technique that relieves asymmetric aging in GPGPU PEs and mitigates the associated timing violations. The proposed solution uses a pseudo-random bit sequence (PRBS) generator that is activated on idle slots of GPGPU execution units. The PRBS circuitry generates dynamic random data patterns that are injected into the GPGPU PEs, thereby avoiding a constant idle state that contributes to asymmetric aging.

The remainder of this paper is structured as follows: Section 2 provides background and discusses prior works. Section 3 investigates the vulnerability of GPGPUs to transistor aging and presents our mitigation approaches and experimental results. Finally, Section 4 concludes our work.

**2. Background and Prior Works**

Transistor aging refers to the deterioration of transistors in digital circuits [5,6], which is caused by charge-carrier trapping in the transistor inversion channel at the dielectric insulator of the transistor gate. The BTI is recognized as the primary mechanism governing transistor aging; it is activated when a constant voltage is applied to the transistor gate, elevating the transistor’s threshold voltage. This increase in threshold voltage increases transistor switching delay and reduces transistor speed. Logical gates that remain in a constant idle state of logical 0 are particularly susceptible to aging because p-type transistors are more prone to BTI than n-type transistors. Asymmetric aging, which means the uneven distribution of performance degradation among transistors in an IC, can lead to severe timing issues, including setup and hold timing violations.

Common approaches to combat transistor aging involve incorporating additional timing margins to mitigate the effects of asymmetric aging. However, such approaches often require complex simulation analyses and can lead to overdesign [7]. Other studies [8–10] have proposed models for predicting aging degradation and have explored various solutions, including reducing clock cycle time, resizing transistors, tuning *V*DD, and power gating. Agrawal et al. [11] proposed predicting circuit failure by using sensors placed at various locations in the silicon die. Additional research [12] explored techniques to analyze digital circuits and detect the most vulnerable gates affected by negative BTI (NBTI) stress. This involves using an aging model with BTI-aware libraries and applying aging-aware timing analysis. Abbas et al. [13] proposed executing anti-aging programs instead of idle tasks during periods of low processor use. Gabbay et al. [6] proposed an aging-aware microarchitecture to minimize the effects of asymmetric aging on execution units, register files, and memory hierarchy in microprocessors while minimizing overhead.

**3. Analyzing How Asymmetric Transistor Aging Affects GPGPU Processing Elements**

This section elucidates how transistor aging affects GPGPU PEs. We experimentally extracted the aging profile of GPGPUs using functional simulations and then used aging models derived from the aging profile to comprehensively analyze the timing. Based on the results, we propose a circuitry to avoid asymmetric aging and thereby mitigate timing violations. Finally, we analyze the effectiveness of this scheme.

The experiments were conducted using the GPGPU simulator [14]. The simulation environment incorporated cycle-level modeling of the RTX 2060 [15] GPGPUs, enabling the execution of computing workloads written in CUDA or OpenCL. To cater to the specific experimental requirements, we modified the simulation platform and implemented the necessary mechanisms for accurate measurements. For benchmarking, we used simulation benchmarks employed in the gpgpu-sim ispass 2009 paper [16]. These benchmarks encompass a diverse range of applications, including neural networks, graph algorithms, and complex mathematical calculations.

The signal probability [6] is a widely used technique to assess the BTI stress profile of logical elements. It quantifies the probability of a signal having a logical value of 1. Specifically, it is determined by the ratio of the time during which a signal remains in the logical 1 state to the total elapsed time. A smaller signal probability corresponds to a more pronounced BTI, resulting in performance degradation and potential failure of ICs over time. Figure 1 shows the activity levels observed in RTX2060 Streaming Multi-Processors for two specific benchmarks: the breadth-first search (BFS) algorithm and a neural network (inference). Activity is measured as the percentage of time the execution unit remains active relative to the total elapsed time. The experimental results indicate that, for the BFS benchmark, the integer execution units within all RTX2060 Streaming Multi-Processors were idle approximately 70% of the time. Conversely, for the NN benchmark, the single-precision floating-point units remained idle more than 85% of the time. These results suggest that the GPGPU PEs may be vulnerable to transistor aging because of their prolonged idle periods. Maintaining an idle state for an extended duration increases exposure to aging effects, potentially impacting the reliability and performance of the PEs.

For this case study, we used an integer arithmetic logic unit (ALU) and a single precision floating point unit (FPU) taken from OpenCore.[[1]](#footnote-1)1 We performed full synthesis, place and route, and timing analysis on these modules in a 28 nm process node. The clock frequency for timing signoff for the FPU and ALU was 164 and 240 MHz, respectively. Our timing analysis used aging-aware library models, as described in Ref. [6]. These models account for the impact of BTI by derating cell delays using NBTI degradation factors derived from the signal probability extracted from the functional simulations illustrated in Figure 1.

The results of our timing analysis, presented in Table 1, reveal that BTI can generate significant timing violations in both GPGPU ALUs and FPUs. When aging is not considered (e.g., for a fresh design) no timing violations occur. However, when aging effects are considered, both the FPU and ALU experience setup and hold timing violations. Although setup violations can be alleviated by reducing the clock frequency, hold violations cannot be mitigated at present.



(a)



(b)

**Fig. 1.** Activity measured in RTX2060 (a) integer execution unit and (b) floating point unit for BFS algorithm and neural network benchmarks, respectively.

**Table 1.** Worst negative slack (WNS) and the number of violated timing paths (NVPs) for FPUs and ALUs for fresh design, aged design, and a design with asymmetric aging avoidance.

|  |
| --- |
| Setup WNS [ps] /NVP |
|  | Fresh | Aged  | Asymmetric aging avoidance |
| FPU | 0/0 | −115/469 | 0/0 |
| ALU | 0/0 | −23/1 | 0/0 |
| Hold WNS [ps] /NVP |
|  | Fresh | Aged | Asymmetric aging avoidance |
| FPU | +4.5/0 | −2/7 | 0/0 |
| ALU | +13/0 | −1/10 | 0/0 |

To address how BTI affects GPGPU PEs, we propose adopting a pseudorandom bit sequence (PRBS) generator activated by a low-frequency clock, as illustrated in Figure 2. This approach is inspired by the technique suggested in Ref. [6] for general-purpose microprocessors. As illustrated in Figure 2, the PRBS data patterns are multiplexed with the functional data path inputs through a designated multiplexer. The use of a PRBS generator creates pseudorandom patterns, which are fed into the GPGPU FPU and ALU, thereby mitigating extended periods of constant stress. The PRBS circuitry is timed using a slow-frequency clock during idle time slots of the FPU and ALU. When the PRBS circuitry is enabled, the input multiplexer selects the PRBS data, which is then injected into the data path of the execution units. The clock frequency for this PRBS generator can be set to a few megahertz or less to minimize any potential dynamic power overhead. This technique provides a practical way to reduce the vulnerability of GPGPU PEs to the BTI, enhancing their resilience and prolonging their operational lifetime. Our timing analysis for the FPU and ALU employing the PRBS asymmetric aging avoidance circuitry, demonstrates the successful elimination of all timing violations, as shown in Table 1.



**Fig. 2.** Asymmetric aging avoidance circuitry based on PRBS generator.

**4. Conclusions**

The emerging deployment of GPGPUs in mission-critical systems establishes stringent requirements for the resilience and reliability of GPGPUs. However, the advent of advanced process nodes has also revealed the susceptibility of ICs to reliability issues, specifically those stemming from transistor aging.

This paper examines how asymmetric transistor aging affects GPGPU execution units. Our experimental analysis shows that GPGPU execution units can be highly susceptible to the BTI due to prolonged periods of idle stress. As a case study, we investigate in this work the NVIDIA RTX 2060 GPGPU using BFS and NN benchmarks, and the results indicate that execution units such as the integer execution unit and the FPU can remain idle for around 70% and 85% of the total time, respectively. These results suggest that such execution units are highly susceptible to asymmetric transistor aging.

These concerns were supported by a detailed timing analysis that combines an aging library model with the aging profile derived from functional simulations. The observed timing violations suggest that GPGPU computational elements can asymmetrically age, resulting in setup and hold timing violations. In addition, we introduced an asymmetric aging avoidance circuitry based on a PRBS generator to mitigate asymmetric transistor aging in GPGPU execution units. A detailed timing analysis indicates that the asymmetric aging avoidance circuitry eliminates the timing violations caused by asymmetric transistor aging.

Further research is warranted in the domain of asymmetric transistor aging in GPGPUs and other computational elements. First, exploring adaptive techniques to dynamically adjust clock frequencies, clock latencies, or resource allocation could be promising avenues for mitigating aging-induced timing violations. Second, investigating novel design methods that integrate fine-grained aging-aware optimizations into the microarchitecture could provide more comprehensive solutions. Finally, extending this study to encompass a wider array of benchmarks and real-world workloads is crucial to establishing the robustness and generalizability of the proposed solution.

**References**

[1]. Kirk, D., and W. Hwu. "GPGPU: General-Purpose Computation on Graphics Hardware." ACM SIGGRAPH, vol. 25, no. 3, 2006, pp. 657-666.

[2] Harris, M., et al. "GPGPUs for General-Purpose Scientific Computing: A Survey." Journal of Supercomputing, vol. 73, no. 1, 2017, pp. 3-50.A. Author, Book title, Editor, *Publisher*, 1990.

[3] Campmany, V., Silva, S., Espinosa, A., Moure, J. C., Vázquez, D., & López, A. M. (2016). GPU-based pedestrian detection for autonomous driving. Procedia Computer Science, 80, 2377-2381.

[4] Yang, M. (2018, July). Avoiding pitfalls when using NVIDIA GPUs for real-time tasks in autonomous systems. In Proceedings of the 30th Euromicro Conference on Real-Time Systems.

[5]. M. A. Alam, H. Kufluoglu, D. Varghese, and S. Mahapatra, “A comprehensive model for PMOS NBTI degradation,” Microelectron. Rel., vol. 47, no. 6, pp. 853–862, Jun. 2007. <https://doi.org/10.1016/j.microrel.2006.10.012>

[6] F. Gabbay, A. Mendelson, Asymmetric aging effect on modern microprocessors, Microelectronics Reliability, Volume 119, 2021, 14090, SSN 0026-2714, <https://doi.org/10.1016/j.microrel.2021.114090>.

[7] S. Ogawa and N. Shiono, “Generalized diffusion-reaction model for the low-field charge build up instability at the Si-SiO2 interface”, Physical Review, 51(7):4218–4230, Feb. 1995.

[8] M. A. Alam, H. Kufluoglu, D. Varghese, and S. Mahapatra, “A comprehensive model for PMOS NBTI degradation,” *Microelectron. Rel.*, vol. 47, no. 6, pp. 853–862, Jun. 2007. <https://doi.org/10.1016/j.microrel.2006.10.012>

[9] S. Bharadwaj, W. Wang, R. Vattikonda, Y. Cao, and S. Vrudhula, “Predictive modeling of the NBTI effect for reliable design,” in *Proc. Custom Integrated Circuits Conf.,* Sep. 2006, pp. 189–192.

[10] W. Wang, V. Reddy, A. T. Krishnan, R. Vattikonda, S. Krishnan, and Y. Cao, “Compact modeling and simulation of circuit reliability for 65 nm CMOS technology,” IEEE Trans. Device Mater. Rel., vol. 7, no. 4, pp. 509–517, Dec. 2007.

[11] M. Agarwal, B. C. Paul, Ming Zhang, and S. Mitra, “Circuit failure prediction and its application to transistor aging”, VLSI Test Symposium, pages 277–286, May 2007.

[12] W. Wang, Z. Wei, S. Yang, and Y. Cao, “An efficient method to identify critical gates under circuit aging,” in Proc. Int. Conf. Comput. Aided Des., Nov. 2007, pp. 735–740.

[13] H. M. Abbas, M. Zwolinski, and B. Halak. Aging Mitigation Techniques for Microprocessors Using Anti-aging Software. Chapter 3, Ageing of Integrated Circuits - Causes, Effects and Mitigation Techniques, Springer, Cham. ISBN 978-3-030-23781-3.

[14] M. Khairy, Z. Shen, T. M. Aamodt, T. G. Rogers. Accel-Sim: An Extensible Simulation Framework for Validated GPU Modeling. In proceedings of the 47th IEEE/ACM International Symposium on Computer Architecture (ISCA), May 29 - June 3, 2020.

[15] https://www.nvidia.com/content/dam/en-zz/Solutions/design-visualization/technologies/turing-architecture/NVIDIA-Turing-Architecture-Whitepaper.pdf

[16] P. Harish and P. J. Narayanan. Accelerating Large Graph Algorithms on the GPU Using CUDA. In HiPC, pages 197–208, 2007.

1. 1 [www.opencores.org](http://www.opencores.org) [↑](#footnote-ref-1)