Process Variation Tolerant 3T1D-Based Cache Architectures *

Xiaoyao Liang, Ramon Canal∗, Gu-Yeon Wei and David Brooks
School of Engineering and Applied Sciences, Harvard University, Cambridge, MA
∗Dept. of Computer Architecture, Universitat Politecnica de Catalunya, Barcelona, Spain
{xliang,guyeon,dbrooks}@eecs.harvard.edu, ∗rcanal@ac.upc.edu

Abstract

Process variations will greatly impact the stability, leakage power consumption, and performance of future microprocessors. These variations are especially detrimental to 6T SRAM (6-transistor static memory) structures and will become critical with continued technology scaling. In this paper, we propose new on-chip memory architectures based on novel 3T1D DRAM (3-transistor, 1-diode dynamic memory) cells. We provide a detailed comparison between 6T and 3T1D designs in the context of a L1 data cache. The effects of physical device variation on a 3T1D cache can be lumped into variation of data retention times. This paper proposes a range of cache refresh and placement schemes that are sensitive to retention time, and we show that most of the retention time variations can be masked by the microarchitecture when using these schemes.

1. Introduction

This paper investigates on-chip memory architectures based on dynamic 3T1D memory cells [4], which can provide speeds equivalent to that of 6T SRAM cells. The paper demonstrates a robust memory design based on 3T1D cells in the context of an important latency-critical component of the processor—the L1 data cache—identified by several researchers as one of the most susceptible architectural components to process variations [3, 1, 5]. We discuss how various sources of process variation can be lumped into a single variable—the retention time of the 3T1D cell—and how characteristics of modern processor architectures can mitigate the costs associated with these dynamic memories. This allows 3T1D-based caches to overcome effects of physical device variations, while offering significant performance, stability, and power advantages over traditional 6T-based designs. 3T1D-based memories are amenable to fine-grained schemes to accommodate designs that suffer large within-die process variations. We studied various data retention schemes—different refreshing and replacement policies—and carried out detailed comparisons in performance and power.

This paper takes several steps in the direction of variation-tolerant memory architecture designs. Specifically, the major contributions are:

- We show how process variations are poised to significantly impede further scaling of memories based on 6T SRAMs.
- Given the transient nature of data flow, we show that dynamic memories can be a good candidate for data storage structures within the processor core. We propose to use a 3T1D-based dynamic memory cell as the base for a new generation of on-chip memory designs.
- We show that the proposed cache structure can tolerate very large process variation with small or even no impact on performance by intelligently choosing the retention schemes. Effects of process variations can be absorbed into a single parameter—data retention time—efficiently addressed by architectural solutions.
- Besides performance benefits, the proposed caches are robust to memory cell stability issues and offer large power savings.

2. Comparison Between 6T Static Memory and 3T1D Dynamic Memory

For decades, the traditional 6T SRAM cell has been the de facto choice for on-chip memory structures within the processor core such as register files and caches. DRAM memory technologies have been used for larger off-chip caches or main memories. These choices have been driven primarily by the high access speed provided by SRAMs and the high density provided by DRAMs. In contrast, data within the processor core is transient while data in lower levels of the memory hierarchy must be stored for much longer time periods. This behooves one to consider using dynamic memories at higher levels of the memory hierarchy (e.g., L1 data cache) if high speed accesses are possible.

As shown in the paper, process variation will negatively impact the speed, stability, and leakage of traditional SRAM designs. Solutions like simply sizing up devices have unacceptable area and power penalties. A desire to continue the scaling trend for on-chip memories in nanoscale technologies drives us to seek out revolutionary memory circuit and architecture solutions.

Because the memory storage node in DRAMs is a capacitor, data is stored temporarily and the memory requires periodic refresh if the data needs to be held for extended time periods. On the other hand, a large fraction of the data consumed in the processor core is transient [6, 2], especially data held in the L1 data cache. As shown in the paper, most cache accesses happen within the initial 6K clock cycles after the data is loaded. As long as the data retention time of dynamic memory can meet system requirements, it is actually a better choice for on-chip memory structures that provide temporary data storage, because the temporal characteristics of the data in the processor may reduce or even eliminate the need for refresh. In light of 6T SRAM scaling concerns, this paper investigates the necessary architectural support that can enable the use of dynamic memories for on-chip cache structures.

Obviously, process variation will also impact the 3T1D memory. For example, if the driving capability of the access transistors is reduced due to the variation, the access time to the cell increases. Figure 1 plots the relationship between memory array access time

* The original paper is: "Process Variation Tolerant 3T1D-Based Cache Architectures," Xiaoyao Liang, Ramon Canal, Gu-Yeon Wei, and David Brooks. MICRO 2007, Dec. 2007. The full paper can be downloaded from http://www.eecs.harvard.edu/~dbrooks/liang_micro07.pdf

and retention time. As retention time increases, degradation—due to leakage—of the stored voltage leads to slower accesses. For a fixed access time target, this effect can otherwise be viewed as decreasing retention time. Weaker-than-designed access transistors have the effect of shifting the access time versus retention time relationship such that cell retention time reduces (from $5.8\mu s$ to $4\mu s$). On the other hand, with stronger devices, retention time can increase. It is important to note that process variations do not necessarily impact operating frequency, which is the case for 6T SRAMs. In a 3T1D DRAM, process variability only causes variations in retention time for a specific target access speed. Moreover, variations in the tag array, decoder, sense amplifier, and other peripheral circuitry in the memory can all be absorbed into the retention time variation. Thus, the impact of process variations on a 3T1D memory can all be lumped into a single variable—the retention time. As detailed in the paper, we propose architectural techniques to tolerate very large retention time variations and, thus, effectively reduce the impact of device process variations in 3T1D-based memories. Also, 3T1D cells do not suffer the cell stability issues previously seen in 6T SRAM cells, because there is no inherent fighting. Except for the finite data retention time, a 3T1D DRAM cell is stable.

In addition to tolerance to process variations, close examination and comparison of 6T SRAM versus 3T1D DRAM cells shows the potential for leakage power savings. Static data storage in a 6T cell results in two leakage paths internal to each cell. In contrast, a 3T1D cell has one leakage path within the cell that eventually degrades the stored data, and once the charge has leaked away, the leakage path is effectively shut off. Hence, the additional power required for refresh is offset by savings in leakage power, which can be a dominant component of the total power for memory arrays in nanoscale technologies.

3. 3T1D-based Cache Architecture

3T1D cells provide many benefits compared to traditional 6T memory cells, but present challenges that must be overcome with system-level support. The major issue associated with 3T1D memories is the limited data retention time and variations of this retention time. Given the relationship between access latency and retention time, the cell can only hold the data for a short time period to ensure the fast-access requirement of on-chip memories. After that time, the cell needs to be refreshed or the data is lost.

This paper proposes two orthogonal categories of techniques to address retention time issues: refresh-based policies and line replacement policies. Refresh policies can be applied at the cache or line-level and replacement policies are extensions to conventional replacement policies (LRU).

Refresh operations require a read and a subsequent writeback to the memory. To avoid large area and power overheads, this paper only considers refresh mechanisms that leverage existing ports in the cache to perform refresh operations. Whenever a refresh is needed, one read and one write port are blocked from normal cache accesses and are used to refresh the data. The refresh mechanism pipelines these read and write accesses on consecutive cycles. However, this approach can introduce a performance penalty, because the refresh operation competes with normal instructions. Moreover, cell-to-cell variations in retention times required by each 3T1D memory cell can complicate this refresh strategy. Fortunately, under-utilization of processor resources in out-of-order machines provide ways to hide refresh operations, and the associated power and performance penalties can be low.

We consider three classes of refresh policies: *full refresh*, where all lines are refreshed, *no refresh*, where lines are allowed to expire

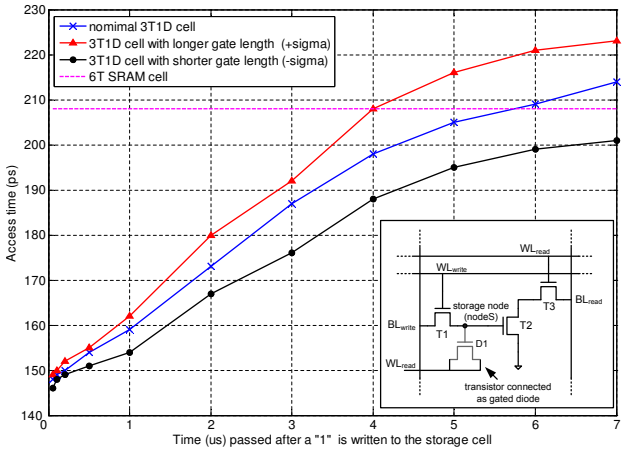

Figure 1. Relationship between access time and retention time for 3T1D cell. Schematic of 3T1D cell (inset).

and L2 cache inclusion is used to provide data, and *partial refresh*, which applies a mixture of the two policies.

We also study four replacement policies: standard LRU replacement, dead-line sensitive placement (DSP), which avoids placement in lines with very low retention time, and two retention-sensitive placement (RSP-LRU and RSP-FIFO) policies which reshuffle cache lines based on variations in the retention time of individual cache lines.

Given the refresh policies and replacement polices, we analyze interesting combinations of the aforementioned policies. The cross-product of the three refresh schemes (no-refresh, partial-refresh, full-refresh) and the four replacement policies (LRU, DSP, RSP-LRU, RSP-FIFO) yields eight viable retention-time strategies to consider. Using a detailed model for variation across a three advanced process technologies, we generated 100 sample chips under severe and typical process variations. To evaluate the efficiency of the schemes, we pick one good chip (representing a chip with process corners leading to very high retention times), one median and one bad chip (very low retention times) in terms of the process variation. We have analyzed the performance of these chips for the eight schemes described above and plot the performance relative to an ideal 6T cache design. Figure 2 summarizes the results. For the bad chip, differences between the schemes are most prominent and we see that RSP-LRU can achieve performance within 3% of an ideal 6T memory (with no variations). The paper discusses the implementation complexity of the schemes and we recommend partial refresh + DSP as being the most complexity-effective scheme.

Two categories of schemes (refresh and placement) have been proposed for one the most important on-chip memory structures – the L1 data cache. By leveraging under-utilization of architectural resources, transient data characteristics, and program behavior, significant performance and power benefits have been demonstrated for 3T1D memories in comparison to 6T SRAM memories. Under *typical* variations, the proposed schemes perform within 2% of the performance assuming ideal 6T memories. Under *severe* variations, where 6T SRAM caches would suffer a 40% reduction in frequency, the schemes presented can perform with less than 4% performance loss and offer inherent extra benefits of reducing leakage power. These promising results suggest that 3T1D memories can replace existing on-chip 6T memories as a comprehensive solution to deal with process variations.

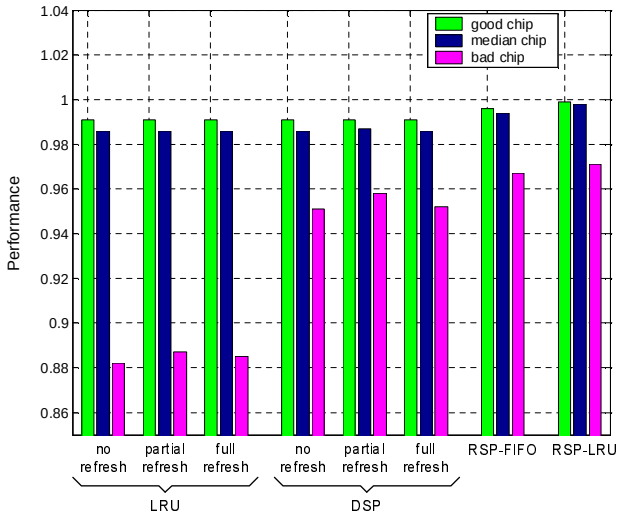

Figure 2. Performance normalized to ideal 6T-based SRAM cache of retention schemes for good, median and bad chips.

Novelty and Industry Relevance

Nanoscale technology scaling offers the promise of continuing transistor density and performance trends. However, the road is fraught with difficulties resulting from increased process variations that limit performance gains and affect stability of key circuit blocks such as on-chip memories. Addressing these problems will require innovation from all levels in the design flow from the devices to the system architecture. This paper investigates using dynamic memory cells with architecture support to enable robust data cache designs tolerant to process variations for future generations of high-performance microprocessors.

Process variation is mainly caused by fluctuations in dopant concentrations and device channel dimensions. Gate length variation can change the effective driving capability of the transistor, as well as the threshold voltage due to the short channel effect. Random dopant variations can also change the threshold voltage of the device. The nature of the semiconductor manufacturing process gives rise to both within-die variations (i.e. device features on one chip can be different) and die-to-die variations (i.e. device features across chips can be different). As technology scales, within-die variations are getting larger, significantly affecting performance and compromising circuit reliability.

On-chip memories consume a significant portion of the overall die space in modern microprocessors given slow off-chip memories and due to the area efficiency and high system performance they offer in exchange for the space and power they consume. On-chip caches traditionally rely on SRAM cells, which have generally scaled well with technology. Unfortunately, stability, performance, and leakage power will become major hurdles for future SRAMs implemented in aggressive nanoscale technologies due to increasing device-to-device mismatch and variations. Circuit imbalance due to mismatch can compromise cell stability (cell flip problem) and exacerbate leakage for traditional SRAM designs. Furthermore, the large number of critical paths and the small number of devices in each path both contribute to increase access time variability. This variability is especially important for higher levels of the memory hierarchy that are more latency critical such as the L1 data cache. In contrast, while lower levels of the memory (e.g., L2 cache) are less sensitive to latency, they remain susceptible to bit flips under process variation. One simple solution to these problems is to slow down scaling of SRAMs at the expense of lower performance and larger area. However, this would effectively mean the end of Moore's Law scaling of transistor density and speed for future processor designs. To address these problems, novel memory circuits and architectures are needed as alternatives to traditional 6T SRAMs, throughout all levels of the memory hierarchy, with different sets of tradeoffs. Due to space limitations, this paper focuses on the L1 data cache.

This paper uses 3T1D memories as the example DRAM cell posited as a viable replacement for SRAM cells. Detailed circuit-level simulations confirm that 3T1D cells offer several advantages in terms of speed and robustness. Moreover, reads are non-destructive such that column multiplexing is possible. While 3T DRAM cells are not new and 3T1D cells have been demonstrated by IBM, this paper is the first to recognize their potential to combat variability concerns via a single parameter—cell retention time—with support via the architecture. Based on promising results using 3T1D cells, it is important to note that other dynamic memory cells such as 1T and 4T cells ought to be studied as well in the future. Single transistor DRAM cells offer even higher levels of density, but their lower access speeds and destructive reads will require additional architectural support to make them viable replacements to the 6T and 3T1D cells.

Besides the L1 data cache, further study is also needed to evaluate the applicability of dynamic memories to other blocks with a microprocessor such as the L1 instruction cache, queues, buffers, register files, etc. Some of these blocks are even closer to the execution core with even lower data longevity requirements. While these blocks are individually small in size, the aggregate die space they collectively consume is significant. Hence, the issue regarding scalability of 6T cells must be addressed. Again, the transient nature of data close to execution core behooves designers to revisit the original decision to rely on static 6T cells for on-chip memories and consider alternatives based on dynamic memory cells.

Microprocessors tolerant to process variations in future nanoscale technologies will be at the forefront of innovation for years to come. This paper proposes novel process variation tolerant on-chip memory architectures based on a 3T1D dynamic memory cell. The 3T1D DRAM cell is an attractive alternative to conventional 6T cells for next-generation on-chip memory designs since they offer better tolerance to process variations that impact performance, cell stability, and leakage power.

References

[1] A. Agarwal, B. C. Paul, H. Mahmoodi, A. Datta, and K. Roy. A process-tolerant cache architecture for improved yield in nanoscale technologies. *IEEE Transactions on Very Large Scale Integration Systems*, 13(1), January 2005.

[2] D. Burger, J. Goodman, and A. Kagi. The declining effectiveness of dynamic caching for general-purpose microprocessors. Technical Report TR-1216, U.W. Madison, Computer Science, 1995.

[3] E. Humenay, D. Tarjan, W. Huang, and K. Skadron. Impact of parameter variations on multicore architectures. In *Workshop on Architectural Support for Gigascale Integration (ASGI-06, held in conjuction with ISCA-33)*, 2006.

[4] W. K. Luk, J. Cai, R. H. Dennard, M. J. Immediato, and S. V. Kosonocky. A 3-transistor DRAM cell with gated diode for enhanced speed and retention time. In *2006 Symposium on VLSI Technology and Circuits*, June 2006.

[5] S. Ozdemir, D. Sinha, G. Memik, J. Adams, and H. Zhou. Yield-aware cache architectures. In *39th IEEE International Symposium on Microarchitecture*, December 2006.

[6] D. A. Wood, M. D. Hill, and R. E. Kessler. A model for estimating trace-sample miss ratios. In *ACM SIGMETRICS*, June 1991.