




Microprocessors and Microsystems

Volume 101, September 2023, 104864

Energy efficiency in multicore shared cache by fault tolerance using a genetic algorithm based block reuse predictor

Avishek Choudhury ^a  , Brototi Mondal ^b , Kolin Paul ^c , Biplab K. Sikdar ^d 

[Show more](#) 

 Share  Cite

<https://doi.org/10.1016/j.micpro.2023.104864> 

[Get rights and content](#) 

Abstract

Aggressive voltage scaling to reduce energy consumption in Multicore causes exponential cell failures in SRAM. Last-level-cache (LLC), the major contender of chip area, exhibits highest sensitivity to low voltage parametric failures and limits energy efficiency. To break the energy barrier, exclusive fault protection mechanism is solicited to ensure on-chip data recovery out of the low voltage failures. This work proposes Cache evolution for energy efficiency by protecting cache blocks from SRAM cell failures due to voltage scaling. A set of coherence and reuse aware in-cache selective replication policies are proposed to ensure on-chip data recovery. Reuse likelihood is predicted through a vector optimization technique using Genetic Algorithm (GA). Reuse aware selective invalidation is employed to balance cache load due to replications. A replication aware replacement policy is also developed that victimizes the lowest reusable cache block. The proposed scheme is evaluated in Multi2Sim 5.0 simulation framework with SPEC CPU benchmark programs. Experimental results claim 43.66% and 38.80% reductions in miss rate and 59.10% and 52.73% reductions in vulnerability for integer and floating point benchmarks respectively.

Energy reduction of 39.21% is observed for integer and 25.73% is observed for floating point benchmarks. Up to 34.78% and 26.98% power reductions in integer and floating point benchmarks are also observed. The minimum supply voltage of 350 mV is achieved at the cost of 7.05% area overhead and 3% performance drop-off.

Introduction

Exponential increase in the number of transistors on chip, in accordance to the Moore's Law, influence power consumption and limits the operational life of the battery operated devices [1]. Supply voltage scaling is maneuvered to reduce power consumption due to quadratic dependency of dynamic/leakage power on supply voltage [2]. However, aggressive voltage reduction to prolong battery life exhibits Process-Voltage-Temperature (PVT) variations that limits Near-Threshold- Computing (NTC) [3]. Memory structures are more susceptible to failure under voltage scaling due to their lower voltage margin [4]. SRAM, with area optimization at nanoscale regime, exhibits highest sensitivity to parametric failures causing permanent faults [5]. Also high energy neutrons from the neutron flux due to cosmic ray strike on earth's atmosphere [6], low energy thermal neutrons from the radioactive ^{10}B boron isotopes in silicon-wafer [7] and alpha particles emitted by the uranium and thorium impurities in packaging material [6] induce soft errors.

Amongst all cache levels, last-level-cache (LLC) is the major contender of chip area. Contemporary processors contain large volume of LLC. Intel's 32 nm Sandy Bridge Core i7-3960X processor contains 15 MB LLC [8]. The 45 MB LLC in Xeon E7 server class processor and even 128 MB LLC in IBM Power 8 processor occupy more than 50% of the die area [9]. Voltage reduction below the minimum supply voltage (VDD_{min}) results in exponential rise of errors in LLC for which exclusive fault protection techniques are employed [10].

Error-correcting-codes (ECCs) are widely used to protect dirty data with added overheads. SECDED code like Hamming Code incurs 12.5% area overhead where 8 bits are required to protect 64-bit cache word [7]. Latency minimization of ECCs results in performance loss [11]. Stronger ECCs incur significant hardware and runtime overheads [4]. Due to ECC overheads, alternative techniques like bit line interleaving [7] and cache scrubbing [12] have been proposed. But multiple word-line grouping increases cache access latency in bit interleaving [7]. On the other hand, the cache scrubbing suffers from tricky scrub interval prediction and inability to correct spatial errors [12]. All these necessitate alternate solutions for fault protection that uses Error Detection Codes (EDCs) for fault detection and techniques like (1) Storing dirty data in reliable cache [13], (2) Cache cleaning by memory write-back [14], (3) Voltage scaling by data redundancy in multiple cache levels [13] and (4)

Additional replication cache for fault protection [15]. But these suffer from restricted voltage reduction, increased memory write-back, increased latency and energy consumption.

On this outset, this work proposes cache evolution, a low latency energy efficient cache system in multicore. A set of coherence and reuse aware selective replication schemes are proposed to find the optimum strategy. Where the state-of-the-art redundancy based techniques [13], [16], [17], [18], [19], [20] use memory access map to track the replicated copies, this work introduces static replication mapping to track the replications without any lookup based approach (by decoding the memory address bits). It reduces the area overhead significantly. Reuse likelihood is estimated through a Genetic Algorithm (GA) based vector optimization technique. The coherence aware selective invalidator is employed to minimize the replication overhead. To boost the underlying fault protection scheme, a reuse aware replacement policy is developed that protects the replicated copy/s against eviction. The key contributions of this work can be summarized as:

- Reuse likelihood prediction of cache block through Genetic Algorithm (GA) based vector optimization technique with the target for implementation in hardware.
- Developing a set of reuse aware cache replication policies with static mapping of cache blocks to minimize the area overhead of using memory access map.
- Introduction of the schemes that realize invalidation of non-reusable blocks to reduce vulnerable time and replication overhead.
- Framing of reuse aware replacement policy to protect the replicated/reserved blocks from evictions.

Section 2 of this paper reveals the related works. The motivation of the proposed work is described in Section 3. Section 4 introduces the details of the proposed energy efficient cache system. Experimental setup to evaluate the performance of the cache system is described in Section 5. The experimental results are reported in Section 6. Section 7 concludes the paper.

Access through your organization

Check access to the full text by signing in through your organization.

Section snippets

Related work

The related works that have been reported so far can be broadly classified either as redundancy based cache fault tolerance techniques or techniques for energy reduction through voltage scaling. ...

Motivation

Both hard and soft errors are reported to be exacerbated under voltage scaling [30], [31], [32], [33]. Hard errors, detected at boot time, are easy to mask. But soft errors being stochastic in nature, are not predictable. Therefore, cache blocks are to be protected in advance to guard against faults, specially soft errors. In the cache system following MOESI protocol, dirty cached copy of a block B in Modified (M)/Owned (O) state is protected at inception (when the block is brought to the cache ...

The fault resilience technique

This section describes the proposed cache system that probes voltage reduction by mitigating cell failures in SRAM. To break the energy barrier, selective protection of vulnerable cache blocks is leveraged based on their coherence state and reuse likelihood. Reuse probability is predicted through a Genetic Algorithm (GA) based vector optimization scheme, designed in hardware. It optimizes the reuse vectors maintained per set for better replication-invalidation-replacement decisions for future ...

Performance evaluation setup

To evaluate the performance of proposed scheme, a multiprocessor system is modelled in Multi2Sim 5.0 simulation framework [45]. SPECrate2017 and SPECspeed2017 integer and floating point benchmark suites of X86 ISA have been used as workloads. Also simulations have been done on SPEC CPU-2006 benchmarks for performance comparison with the

existing fault resilient approaches. All the SPEC CPU 2017 benchmarks are used to build the workloads. Simulation results for the 14 (7 integer and 7 floating ...

Experimental results

This section reports the simulation results to evaluate the performance of proposed cache system within a range of voltage variations. The outcomes of experimentation are summarized in the following sub-sections. ...

Conclusion

Energy efficient LLC evolution is proposed in this work for multi-core. Five different replication-invalidation-replacement techniques are proposed with a genetic algorithm based reuse prediction. All the schemes are evaluated under voltage variations on several performance metrics. *One clean replication in local cache in single trial and one dirty replication in remote bank/s in double trials* is found to be the most effective. Experimental results claim 43.66% and 38.80% reductions in miss ...

Avishek Choudhury received the Bachelor of Science (Honours) in Computer Science from North Bengal University, West Bengal, India in 2006, the M.C.A. and M.Tech in Computer Science and Engineering degrees from WB University of Technology, West Bengal, India in 2009 and 2013 respectively and pursuing Ph.D. in Computer Science and Technology from IIST Shibpur, India. Presently, he is working as Assistant Professor in the Department of Computer Science at New Alipore College, affiliated to the ...

...

...

[Recommended articles](#)

References (59)

RobertsDavid *et al.*

[On-chip cache device scaling limits and effective fault repair techniques in future nanoscale technology](#)

Microprocess. Microsyst. (2008)

DaiHongjun *et al.*

Security enhancement of cloud servers with a redundancy-based fault-tolerant cache structure

Future Gener. Comput. Syst. (2015)

FerrerónAlexandra *et al.*

A fault-tolerant last level cache for CMPs operating at ultra-low voltage

J. Parallel Distrib. Comput. (2019)

FarbehHamed *et al.*

CLEAR: Cache lines error accumulation reduction by exploiting invisible accesses

Microelectron. J. (2019)

AlameldeenAlaa R. *et al.*

Energy-efficient cache design using variable-strength error-correcting codes

ACM SIGARCH Comput. Archit. News (2011)

DhimanGaurav *et al.*

Dynamic voltage frequency scaling for multi-tasking systems using online learning

Shekhar Borkar, Tanay Karnik, Siva Narendra, Jim Tschanz, Ali Keshavarzi, Vivek De, Parameter variations and impact on...

MittalSparsh

A survey of architectural techniques for near-threshold computing

ACM J. Emerg. Technol. Comput. Syst. (JETC) (2015)

RobertC Baumann

Radiation-induced soft errors in advanced semiconductor technologies

IEEE Trans. Device Mater. Reliab. (2005)

SlaymanCharles W.

Cache and memory error detection, correction, and reduction techniques for terrestrial servers and workstations

IEEE Trans. Device Mater. Reliab. (2005)



View more references

Cited by (0)



Avishek Choudhury received the Bachelor of Science (Honours) in Computer Science from North Bengal University, West Bengal, India in 2006, the M.C.A. and M.Tech in Computer Science and Engineering degrees from WB University of Technology, West Bengal, India in 2009 and 2013 respectively and pursuing Ph.D. in Computer Science and Technology from IIST Shibpur, India. Presently, he is working as Assistant Professor in the Department of Computer Science at New Alipore College, affiliated to the University of Calcutta. Before that, he worked as UGC research fellow at the CVPR Unit, ISI Kolkata. His research interests include fault tolerant computer architecture design and machine learning.



Brototi Mondal received the Bachelor of Technology in Information Technology from WB University of Technology, India in 2010 and the M.Tech in Computer Science and Engineering from WB University of Technology, India in 2012. Presently, she is working as Assistant Professor in the Department of Computer Science at Sammilani Mahavidyalaya, affiliated to the University of Calcutta. Before that, she worked as Assistant Professor in the Department of Computer Science and Engineering at the SKFGI, WB, India. Her research interests include fault tolerance, computer architecture, machine learning and cloud computing.



Kolin Paul is a Professor in the Department of Computer Science and Engineering at IIT Delhi India. He received his B.E. degree in Electronics and Telecommunication Engineering from NIT Silchar in 1992, ME from Jadavpur University in 1995 and Ph.D. in Computer Science in 2002 from BE College (DU), Shibpore. During 2002-3 he did his post doctoral studies at Colorado State University, Fort Collins, USA. He has previously worked at IBM Software Labs. His last appointment was as a Lecturer in the Department of Computer Science at the University of Bristol, UK. He has also held a Visiting Position at KTH, Stockholm. His research interests are in understanding high performance architectures and compilation systems. In particular he works in the area of Adaptive/Reconfigurable Computing trying to understand its use and implications in embedded systems. He is also involved in the design of systems for affordable healthcare.



Biplab K Sikdar received the Bachelor of Science (honours) in physics from Presidency College, Calcutta University, Calcutta, India in 1985, the B.Tech and M.Tech degrees in computer science and engineering from Calcutta University in 1988 and 1990 respectively, and the Ph.D. in engineering from Bengal Engineering College (DU), West Bengal, India in 2003. He was with the Faculty of Computer Science and Engineering, North Eastern Regional Institute of Science and Technology, Itanagar, India from 1991 to 1992 and in the University of North Bengal, Siliguri, India from 1992 to 1997. He is serving Indian Institute of Engineering Science and Technology since 1997 and presently, he is a Professor in the Department of Computer Science and Technology. His research interests include digital system design and test, computer architecture and in-memory computation. He has been working on the development of hierarchical cellular automata for VLSI design and test.

[View full text](#)

© 2023 Elsevier B.V. All rights reserved.



All content on this site: Copyright © 2025 Elsevier B.V., its licensors, and contributors. All rights are reserved, including those for text and data mining, AI training, and similar technologies. For all open access content, the relevant licensing terms apply.

