Microelectronics Reliability 55 (2015) 2439–2452

Contents lists available at ScienceDirect

Microelectronics Reliability

journal homepage: www.elsevier.com/locate/mr

Design space exploration of non-uniform cache access for soft-error vulnerability mitigation

Mohammad Maghsoudloo, Hamid R. Zarandi *

Department of Computer Engineering and Information Technology, Amirkabir University of Technology (Tehran Polytechnic), Hafez Ave., Tehran, Iran

ARTICLE INFO

Article history:
Received 16 November 2014
Received in revised form 31 July 2015
Accepted 31 July 2015
Available online 14 August 2015

Keywords:
Soft error
Design space exploration
Many-core processors
Non-uniform cache access
Temporal vulnerability factor

ABSTRACT

In this paper, the design space exploration problem is concerned with finding the best composition of different Non-Uniform Cache Access (NUCA) specifications in many-core processors. The single-objective and multi-objective exploration problems are intended to meet the desired level of reliability without violating the performance and energy constraints. The main objective is to find the best choice for each cache specification which can minimize the vulnerability of L1 and L2 caches in NUCA architectures. The design space consists of 72 implementations, made up of combinations of different structures in the current NUCA specifications (cache organization, write policy, coherence protocol, inclusiveness, replacement policy, and network topology). Moreover, the effects of design implementations on reliability (as the main objective), performance, cache energy consumption, and interconnection traffic (as the constraints) have been investigated.

1. Introduction

The shrinking of process technologies enables many cores and large caches to be incorporated into tiled architectures where they will be physically distributed throughout the chip [1]. In this architecture, caches are typically partitioned into many banks and are organized in a non-uniform architecture [2]. Each cache bank is located within a tile, and connected to other banks via a mesh-based model of interconnection network [3]. Non-Uniform Cache Access (NUCA) alleviates the issue of the on-chip wire delay of integrated caches by embedding a network in the cache [4]. Moreover, NUCAs increase scalability, and offer greater performance stability than conventional uniform cache architectures in many-core eras. In general terms, NUCAs improve performance per watt of the current many-core processors compared to uniform architectures [1,2].

Despite recent trend on integration of memory hierarchies onto the processing die to improve performance, the reliability of processors has been significantly affected by soft errors as cache memory, composed of SRAM bits, is significantly susceptible to particles strike. This issue gets more serious with the fact that 60% of chip area is occupied by different levels of caches, which are exposed to soft errors [5–9]. This problem has become even more acute with the increasing likelihood of single-event Multi-Bit Upsets (MBUs) in current technology [10]. Although, lightweight detection mechanisms (such as using parity bits and SEC-DED) can inexpensively solve the issue of detecting multiple bit flips, however, the ability of error correction of MBUs cannot be easily provided [6]. The inefficiency of previous correction (recovery) mechanisms in tackling MBUs for dirty data items has been shown previously [6]. Enhancing the capability of previous correction techniques (for covering MBUs) poses significant performance, energy, and area overheads [5]. In current manufacturing technologies, increasing importance of reliability makes it worthy of consideration in design and optimization processes.

During design of these complex many-core chips, many platform parameters have to be tuned. This is done in order to maximize platform main objective while minimizing non-functional costs or secondary objectives. This tuning phase is called Design Space Exploration (DSE) [11]. This process can be formalized as a multi-objective optimization problem where non-commensurable objectives have to be maximized (or minimized) [12]. Specifically for NUCAs, the processor architect attempts to find the optimal cache memory hierarchy given a set of representative applications expected to run on the system. Traditionally, DSE techniques improve cache performance, in terms of miss reduction, at the expense of energy [13–17]. In this approach, all possible cache configurations are simulated to find the optimal solution. Finally, cache parameters are tuned based on performance and energy results. However, the importance of reliability in the DSE has not yet been treated in the literature.

In this paper, the design space exploration problem is concerned with finding the best composition of different NUCA specifications that meet the desired level of reliability without violating the constraints. The main objective is to find the best structure (parameters) for each cache platform (specification) which can minimize the vulnerability of L1 and L2 caches. Temporal Vulnerability Factor (TVF), which

* Corresponding author.
E-mail addresses: m.maghsoudloo@aut.ac.ir (M. Maghsoudloo), h_zarandi@aut.ac.ir (H.R. Zarandi).

http://dx.doi.org/10.1016/j.microrel.2015.07.049
0026-2714/

has been previously defined and used [8], is selected to estimate and compare the vulnerability of different design decisions on reliability. Each design implementation is made up of a composition of the different structures in the following cache specifications: cache organization, write policy, coherence protocol, inclusiveness, and replacement policy. Due to the correlation between the effects of network topology and some cache specifications on the design elements, two different structures of interconnection network topologies are also studied. Moreover, the exploration problem is studied for single and multi-objective perspectives. In the single-objective viewpoint, the effects of design implementations on reliability (as the main objective), performance, and cache energy consumption (as the constraints) are investigated. In the multi-objective formulation, NUCA architectures are configured to simultaneously meet varied performance requirements, and energy budgets, accompanied with vulnerability mitigation.

In our experiments, 72 different configurations are implemented. Cache configurations are made by means of varying parameters of one specification and keeping the parameters of other specifications constant. In order to evaluate the design space, a functional simulation infra-structure, SIMICS [18], is used with an extended version of Multi-facet GEMS [19]. Moreover, the behaviors of 6 programs in the category of SPLASH-2 benchmarks suite [20] and network intensive workloads have been studied on the simulation environment [21,22]. In the case of single-objective DSE, the performance, and energy consumptions constraints are arbitrarily adjusted at three levels. When the performance and cache energy consumption values are tight to satisfy the strictest level of constraint, L1 cache normalized TVF and L2 cache normalized TVF are not less than 81% and 68% of the maximum, respectively. Concerning the midlevel of constraints, L1 cache normalized TVF is decreased to 72% of the maximum, however, L2 cache TVF cannot be decreased any more. When the low-severity of constraints is desired, the L1 and L2 cache normalized TVF are significantly decreased (down to the minimum value among possible configurations). Furthermore, multi-objective DSE is done with respect to three tradeoffs in both L1 and L2 caches. Based on the proximity of their results possible configurations can be classified into four groups. In addition to considering the effects of using different configurations on three tradeoffs, the consistency of these effects on the vulnerability of L1 and L2 caches are investigated. It is checked that changing the parameters of a specification for L1 vulnerability mitigation either leads to improving the L2 vulnerability. Finally, the results of single- and multi-objective DSEs are exploited to investigate the efficacy of specifications on the design elements. This analysis can help identify the most effective specification to improve different design elements.

The rest of paper is organized as follow: Section 2 describes the background and terminologies. The DSE results and discussion are explained in Section 3. Section 4 shows the result analysis. Finally, Section 5 concludes the paper.

2. Preliminaries

2.1. Tile architecture and NUCA Specification

The Tile architectures have their origins in the RAW research project developed at MIT and later commercialized by Tilera [23]. The key differentiating points of the tiled architecture are the on-chip interconnection network and the L2 cache architecture. The chip is structured as a 2D array of tiles, contains a processor core with dedicated L1 caches, a slice of the L2 cache, a slice of the directory and network interfaces connecting the corresponding tile to the Interconnection network. The L2 cache can be either a part of a shared L2 cache or a private L2 for the local tile. In case of a shared L2, cache blocks are address-interleaved among the L2 slices. On the other hand, private L2s are accessed just by its own L2 tile [2,23]. Also depending on the cache organization, the tile includes structures to support distributed cache coherence protocol. In the case of using private organization, coherence must be kept among all sharers at L2 cache. However, L1 caches in shared organization need protocols to be coherent [24].

2.2. Technical approach

The rationale behind the design space exploration problem, tuned by this paper, is to compare and rank different non-uniform cache architectures. DSE is traditionally used to achieve optimal cache structure [31–33]. In this methodology when the design space is not too large, all possible cache configurations are simulated using a cache simulator to find the optimal solution. Cache configurations are made via varying parameters of one specification and keeping the parameters of other specifications constant. Afterwards, the configurations, which cannot meet the constraints, are eliminated from the set of possible configurations. Finally, the possible cache configurations are sorted concerning the main design objective [13–17]. To have trade-off between two or more than two specifications, a multi-objective DSE can be formulized. In this case, all Pareto-optimal configurations should be investigated concerning the requirements of two or more than two objectives.

Regarding Table 1, the design space of the above DSE problem is obtained by varying popular existing structures (parameters) for NUCA cache specifications and network topology [25,26]: cache organization (shared or private), cache write policy (write-back or write-through), coherence protocol (MOESI or MESIF), inclusiveness policy (exclusive, non-inclusive, or inclusive), replacement policy (two versions of pseudo LRU), and interconnection network topology (bus, or 2D-Mesh). These parameters are chosen based on the common characteristics of current modern microprocessors [25,26]. While the effects of two important specifications (cache line size, and degree of associativity) on the design objectives have been investigated by [8], their effects are not intended in the main analysis of this paper. However, in [8], their effects have been analyzed with regard to single-threaded workloads and simulation environment. In Section 4.3, an analysis is made to investigate that multi-threaded workloads and simulation environment can either change the effects of these two specifications on cache vulnerability or not. The objective of our approach is to obtain a set of optimal cache configurations with respect to L1 and L2 cache vulnerability for a given performance and energy consumption levels of both L1 and L2.

During experimental simulations, the parameters of other components of inter-core and intra-core specifications were kept constant (such as number and type of the processing cores). The specification of base system is presented in Table 2. While the scope of this paper is limited to DSE of NUCA with a focus on reliability, the parameters of specifications of RAW-based many-core processor are kept constant.

2.3. Simulation environment

In order to evaluate the design decisions, a functional simulation infra-structure, SIMICS [18], is used accompanied with Multi-facet

Table 1
Cache specifications and parameters.

Cache organization	Write policy	Coherence protocol	Inclusiveness policy	Replacement policy	Interconnection network topology
Shared	Write-back	MOESI	Exclusive	Pseudo LRU based on tree bits (TLRU)	Bus
Private	Write-through	MESIF	Non-inclusive	Pseudo LRU based on MRU bits (MLRU)	2D-mesh
			Inclusive		