Integration, the VLSI Journal xxx (2018) 1–22

Contents lists available at ScienceDirect

Integration, the VLSI Journal

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

Reliability-aware application mapping onto mesh based Network-on-Chip

Navonil Chatterjee *, Priyajit Mukherjee, Santanu Chattopadhyay

Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology, Kharagpur, WB 721302, India

ARTICLE INFO

Keywords:
Network-on-Chip
Application mapping
Particle Swarm Optimization
Mixed Integer Linear Programming
Reliability
Average packet delay

ABSTRACT

In Network-on-Chip based multi-core systems, application mapping is a critical issue as it affects the overall system performance in terms of average packet delay (APD) and system reliability. As the number of Intellectual Property (IP) cores integrated into a single chip increase, heat dissipation becomes a major issue. Due to non-uniform heat spreading, thermal hotspots are formed which reduces the system reliability. In this work, we have proposed a reliability model which in turn considers thermal effect due to computation and communication entities present in the network. Based on the reliability model, we have proposed a Mixed Integer Linear Programming (MILP) formulation of the mapping problem for improving system reliability with a constraint on the APD of the network. To address large sized applications, we have used a Particle Swarm Optimization (PSO) based mapping method. The basic PSO has been augmented with a constructive heuristic to improve the system reliability further. Experiments have been conducted on applications belonging to SPLASH-2 and PARSEC benchmark suites, which shows significant improvement in system reliability, compared to other works mentioned in the literature.

1. Introduction

Advancement in VLSI technology and scaling of transistors to nanometer range have made it possible to integrate billions of transistors on a single chip. This has helped the system designers to embed a large number of Intellectual Property (IP) cores, Memory Units (MUs) and Processing Elements (PEs) onto a single chip, resulting in a System-on-Chip (SoC). SoC architecture with a large number of cores requires high performance of on-chip communication medium between the components. Major challenges faced by traditional bus-based SoC include shared inter-core communication, high latency and increased power consumption. To overcome the drawbacks of SoC, Network-on-Chip (NoC) has been proposed as a viable solution. However, aggressive technology scaling and increasing design complexity of NoC based MPSoC have made the chip vulnerable to faults. These faults are either permanent or transient in nature. Permanent faults, as the name suggests, are nonrecoverable device defects. These faults may be introduced by manufacturing defects or device wear-out caused by Negative Bias Temperature Instability (NBTI) [1], Electro-Migration and Oxide Breakdown [2]. On the other hand, transient faults are temporary random faults which occur for a very short interval of time (order of millisecond). Transient faults may occur due to crosstalk [3], effects of alpha particles and cosmic radiation [4], etc. These faults negatively affects the system reliability. Thus, a major challenge in this domain is that of designing a NoC based MPSoC, that can enhance the system reliability.

NoC interconnection framework primarily consists of routers and links. A router may be associated with a core through a network interface (NI) module. NI acts as an interface between the computation and communication units. At the source, NI converts the data from the IP cores into packets. Whereas, at the destination, NI joins all such packets to get back the message and delivers to the IP core [5]. Routers route these packets from source to destination nodes utilizing the underlying network topology and routing algorithm. Each router is connected to its neighbours by point-to-point communication links, which constitute the NoC topology. Some of the well known topologies of NoC are Mesh, Ring, Torus, Star, Butterfly, Fat Tree etc. [6]. The desired NoC architecture should have the following features: large bisection bandwidth, low node degree, small diameter and low average distance [7,8]. Mesh has turned out to be the most popular among all the topologies [9]. It has large bisection bandwidth, simple node connection and low complexity routing algorithm, with no long links and ease of physical implementation. Due to the aforementioned reasons, mesh topology has been selected in this work.

An application is represented as a collection of tasks which communicate among themselves to realize a specific function. IP cores

* Corresponding author.
E-mail addresses: navonil@iitkgp.ac.in (N. Chatterjee), priyajit@iitkgp.ac.in (P. Mukherjee), santanu@ece.iitkgp.ernet.in (S. Chattopadhyay).

https://doi.org/10.1016/j.vlsi.2018.02.002
Received 23 June 2017; Received in revised form 15 December 2017; Accepted 1 February 2018
Available online XXX

present in the design library are used for executing specific tasks. First step in NoC design is the allocation of tasks to appropriate IP cores. The selected cores form a core graph, where vertices correspond to cores and each edge corresponds to communication between a pair of cores, with communication bandwidth as its label. Allocation of IP cores to different routers in the network platform, also known as application mapping, is an important issue as the overall system performance depends on it [10]. Application mapping has been a very well researched domain. Many NoC mapping strategies are available in the literature. The major challenges regarding application mapping in NoC are - (1) minimization of average packet delay, (2) improving system reliability (3) reduction of peak temperature, (4) reduction of network contention, (5) optimization of traffic to reduce congestion and so on. In this work, we have focused on the maximization of system reliability, while minimizing average packet delay (APD). To obtain an optimized mapping in terms of APD for a given application, a contiguous processor allocation is desirable, such that highly communicating cores are mapped close to one another. This helps to reduce the APD of the system. However, this may lead to the high power consuming cores being placed close to each other, resulting in thermal hotspots. This negatively affects the performance of NoC and reduces system reliability. As the failure rate of cores is dependent on their operational temperatures, thermal-aware task mapping schemes may be employed to improve system reliability. However, such schemes act only upon the temperature and do not always guarantee an optimized result in terms of system reliability. So, an effective reliability model, depending on the temperature of individual cores, is necessary to solve the mapping problem. However, an optimization approach targeting system reliability may significantly elevate communication overhead, resulting in increased system latency. Therefore, to obtain an optimized mapping solution in terms of reliability and APD, a trade-off between the two is required. To solve this problem, any search based algorithm could be developed which would do a search space exploration for the same. However, with the increase in the number of cores in the application core graph, the search space increases exponentially. Exhaustively searching the solution space becomes almost impossible. Several optimization techniques based on heuristics, meta-heuristics and evolutionary approaches have been developed to explore the same. The techniques presented in NMAP [11], Kernighan-Lin (KL) [12], Simulated Annealing (SA) [13], Particle Swarm Optimization (PSO) [14], Genetic Algorithm (GA) [15] etc, provide good search space exploration. In Ref. [16], the authors have shown that PSO performs better compared to GA, as it takes lesser computational effort for convergence. Also, the SA algorithms require higher number of iterations to converge than PSO [17]. It has also been observed from Ref. [14], that PSO based application mapping produces better results than the greedy heuristic, such as, NMAP [11] and KL [12]. This motivates us to select PSO for solving the problem of reliability-aware application mapping.

In our previous work [18], we have solved the reliability-aware mapping problem via a constructive heuristic algorithm. The reliability model presented in the paper considered tile based NoC, where each tile comprised of a core and a router. However, while calculating system reliability, only the temperature of the cores had been considered. The works presented in Refs. [19,20], have shown that routers in NoC have a comparable thermal impact as that of the processors and contribute significantly to the overall chip temperature. This is because the routers have higher switching activity in smaller chip area compared to the processors. Thus, in this work, we have considered both core and router power values while calculating system reliability. We have also presented a thermal model based on HotSpot [21], that helps in calculating the temperature of the chip. The proposed reliability model has been incorporated in the mapping technique. The objective function used in this paper includes reliability and APD. A weight factor has been incorporated to control the amount of emphasis put on these two objectives. This gives the flexibility to the designer to tune the system, according to the requirements, i.e. reliability or APD. We have used a PSO based mapping algorithm for core to router mapping, which gives better results in terms of reliability, compared to the constructive heuristic approach presented in Ref. [18]. To improve the quality of solution further, we have augmented the PSO based technique with the constructive heuristic method presented in Ref. [18]. It should be noted that in this paper, we have focused on the application mapping problem which directly influences the temperature of the chip. Therefore, we have adopted the temperature dependent failure rate calculation procedure presented in Ref. [22], while modeling the reliability of the NoC based MPSoC. Other factors affecting the system reliability does not fall under the purview of this work.

The novel contributions of this paper are as follows:

- We have proposed a reliability model taking into consideration the temperature of both cores and routers. Also, we have modeled the NoC, considering 2D mesh topology, as a series system, where a drop in reliability of any core or router would reduce the reliability of the entire system. From the proposed reliability formulation, we have also derived the Mean Time To Failure (MTTF) of the system.
- We have developed a compact thermal model based on HotSpot [21], which determines the temperature profile of the chip from the power consumption of cores and routers present in it.
- We have formulated a Mixed Integer Linear Program (MILP) to map the given application (in terms of core graph) onto a Mesh based NoC, which aims at maximizing the reliability of the system while minimizing the average packet delay.
- We have proposed a PSO based mapping strategy, which incorporates both the reliability and APD in its optimization function. Further, we have augmented the basic PSO algorithm with constructive heuristic [18] to generate better initial particles, which improves the search space exploration.

The rest of the paper is organized as follows. In Section 2 we have presented the literature review. Section 3 presents the proposed reliability model. Section 4 contains analysis of the Mean Time to Failure (MTTF) of the NoC. Section 5 presents the problem definition which describes the notations of application mapping and the metrics used for system evaluation along with the objective function. A brief description of the selected thermal model has been presented in Section 6. A Mixed Integer Linear Programming (MILP) for reliability-aware application mapping has been presented in Section 7. Section 8 presents the PSO based reliability-aware mapping algorithm. A detailed description of the experimental setup and associated results have been discussed in Section 9. In Section 10, conclusions and future works have been presented.

2. Literature review

The works presented in the application mapping domain can broadly be classified into two parts - Static Mapping and Dynamic Mapping. In this work, we have focused on mapping which could be implemented at the design time (i.e. static mapping) as it provides better search space exploration and generates a good quality mapping solution. Works related to reliability awareness of the system have also been presented. We have organized this section into two subsections consisting of works with or without reliability awareness.

2.1. Application mapping techniques without reliability awareness

An Integer Linear Programming (ILP) based application mapping technique has been presented in Ref. [23]. The method consists of a unified mapping approach that aims at solving different problems, which includes energy efficient application mapping, operating voltage assignment and routing. Authors in Ref. [24] have proposed an ILP formulation for contention aware application mapping in a tile