This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2015.2422843, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems

An Efficient Application Mapping Approach for the Co-Optimization of Reliability, Energy and Performance in Reconfigurable NoC Architectures

Chen Wu, Chenchen Deng, Leibo Liu*, Jie Han, Jiqiang Chen, Shouyi Yin, Shaojun Wei

***Abstract*—In this paper, an efficient application mapping approach is proposed for the co-optimization of reliability, communication energy and performance in network-on-chip (NoC) based reconfigurable architectures. A cost model for the co-optimization of reliability, communication energy and performance (CoREP) is developed to evaluate the overall cost of a mapping. In this model, communication energy and latency (as a measure of performance) are first considered in energy latency product (ELP), and then ELP is co-optimized with reliability by a weight parameter that defines the optimization priority. Both the transient and intermittent errors in NoC are modeled in CoREP. Based on CoREP, a mapping approach, referred to as priority and ratio oriented branch and bound (PRBB), is proposed to derive the best mapping by enumerating all the candidate mappings organized in a search tree. Two techniques, branch node priority recognition and partial cost ratio utilization, are adopted to improve the search efficiency. Experimental results show that the proposed approach achieves significant improvements in reliability, energy and performance. Compared with the state-of-the-art methods in the same scope, the proposed approach has following distinctive advantages: 1) CoREP is highly flexible to address various NoC topologies and routing algorithms while others are limited to some specific topologies and/or routing algorithms; 2) General quantitative evaluation for reliability, energy and performance are made respectively before integrated into unified cost model in general context while other similar models only touch upon two of them; 3) CoREP based PRBB attains a competitive processing speed, which is faster than other mapping approaches.**

***Index Terms*—Energy, Latency, Reliability, Network-on-Chip (NoC), Mapping Algorithm.**

I. Introduction

With both the flexibility of general purpose processors (GPPs) and the efficiency of application specific integrated circuits (ASICs), reconfigurable architectures have proven their advantages in various application domains [1]. As a promising interconnect infrastructure, network-on-chip (NoC) has a significant impact on the reliability, energy and performance of communications in reconfigurable systems. The reliability of NoC is thus of great importance, because it may cause the failure of the whole system [2][3]. However, the vulnerability of NoCs to factors such as crosstalk [4], electromagnetic interference (EMI) [5], and radiation [6] makes reliable communication very challenging. The communication energy, which accounts for more than 28% of the total energy [7][8] in the NoC, is also significant. At the same time, performance, in terms of latency and throughput, is a critical design parameter as well [9]. Moreover, in recent GPPs, such as the future payload data processing cores for science, earth and telecommunication missions identified by the European and American space agencies [10][11][12], the communication infrastructure is required to have high reliability, low power consumption, and high performance. During an application mapping, it therefore becomes crucial to simultaneously optimize reliability, communication energy and performance of the NoC in a reconfigurable system.

Effort has recently been made to optimize reliability in the mapping procedure [6]. Work has also been done to reduce energy [13][14] or latency [15][16] when searching for an optimal mapping. In [17], the task mapping approach optimizes communication energy while faults in the NoC are tolerated. Mapping approaches for optimizing energy and latency have also been proposed [5]. In [18], reliability and energy are both considered in finding the best mapping for multiple applications. The performance overhead is minimized by simply mapping on a rectangular area. In [19], a quantitative model of energy and reliability is proposed and the performance is qualitatively considered by using bandwidth constraints. In [20] and [21], energy, reliability, and throughput are all considered during the mapping procedure. Although an improvement in reliability, energy, and performance can be obtained by these approaches, they have the following major disadvantages: 1) These approaches are limited to a specific NoC topology and/or a routing algorithm. The models in [18] and [19] consider to map on a rectangle or to make the bounding box of the source-destination pair closer to a square, which may not work for topologies other than a mesh. 2) Although the approaches proposed in [20] and [21] are topology independent, their model is quite simple that only energy/time is quantitatively modeled while others are qualitatively considered. In other approaches, reliability and energy are quantitatively evaluated, and performance, which is also as important as these two measures, cannot be quantitatively evaluated. 3) The computational complexity of the mapping approaches is rather high due to the lack of appropriate models and efficient mapping algorithms.

In addition, the integration of these three metrics cannot be

Chen Wu, Chenchen Deng Leibo Liu, Jiqiang Chen, Shouyi Yin, and Shaojun Wei are with the Institute of Microelectronics and The National Lab for Information Science and Technology, Tsinghua University, Beijing, 100084, China (*E-mail of correspond author: liulb@tsinghua.edu.cn).

Jie Han is with the ECE Department, University of Alberta Edmonton, Canada, T6G, 2V4.

This work is supported in part by the China National High Technologies Research Program (No. 2012AA012701) and the project from State Grid Cooperation of China (No. SGRI-WD-71-13-014/008/010/011).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2015.2422843, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems

achieved by simply combining the off-the-shelf approaches. Each of them is set within a specific context, and it is almost impossible to integrate three formulations in different works together to gain a general model. Therefore, reliability, energy and performance should be evaluated into a unified model. Flexibility is of great importance when proposing these models. In NoC-based reconfigurable architectures, the configuration of processing elements (PE) functions and the interconnections are dynamically changed to obtain high flexibility. Thus the topology and routing algorithm for the NoC differ from one to another in various application scenarios. Therefore, it is mandatory that an application mapping approach can handle diverse topologies and routing algorithms to satisfy the requirement of reconfigurable systems. Therefore, each of these three models is required to be not only accurate estimation for each metric but also not limited to a specific NoC topology and/or routing algorithm. In addition, dynamic reconfiguration of a reconfigurable architecture requires the speed of the application mapping to be maximized; thus it is important to minimize the computational overhead of a mapping approach.

To address these issues, an efficient application mapping approach is proposed for the co-optimization of reliability, communication energy and performance (CoREP) in this paper. Communication energy and latency are first combined by energy latency product (ELP) and then ELP is combined with reliability with a weight parameter. CoREP is designed to be highly flexible to handle various NoC topologies and routing algorithms. Furthermore, reliability, communication energy and latency are optimized in CoREP simultaneously. Based on CoREP, a mapping approach, referred to as priority and ratio branch and bound (PRBB), is further proposed to find the best mapping pattern. The branch node priority recognition and partial cost ratio utilization techniques are adopted to reduce computational overhead.

II. Model Analysis

A. Background

The application used in this study is presented as an application characteristic graph ($APCG$) $G(C, A)$ [2]. $G(C, A)$ is a directed graph, where each vertex $c_i \in C$ represents an intellectual property (IP) core, each edge $a_{ij} \in A$ represents the communication between c_i and c_j, and the weight of each edge V_{ij} indicates the communication volume on edge a_{ij}. The architecture of an NoC is also represented as a directed graph. In the graph, each vertex denotes a node, which includes a PE and a router, whereas each edge indicates a link connecting the nodes. The links are bidirectional, whose total number is defined as N; for example, for a 4×4 mesh topology, $N = 48$.

The reliability of an NoC-based reconfigurable architecture can be affected by the soft and/or hard errors in PEs, routers and links. To address the problem that PEs are faulty (by soft or hard errors) or routers and links are affected by hard errors (or permanent errors), redundant PEs are made on the chip. When a fault occurs, these spare components can be used to replace the faulty ones to tolerate the fault. [19][22]. After the replacement by spare components, the topology and routing algorithm of the NoC change. In this way, the application is then required to remap onto a reconfigured NoC. This is the problem addressed in this paper that the mapping approach is highly flexible which can be applied to reconfigurable NoC-based architecture with various topologies and/or routing algorithms. Therefore, the proposed model copes with the rest cases of errors which are soft errors for routers and links (both transient and intermittent errors). In the following discussion, the faults are referred to soft faults. For the fault in a router, it may also affect some links connected to the faulty router. For example, the faults in the routing computation module will affect all the output ports of the router. Therefore, the assumption for the worst case are made in this manuscript for simplicity and similar assumption is also made in [23]. This means that faults in any part of a router are considered as the case that faults occur in all the links which are connected to the faulty router. Hence, all the faults for both links and routers are classified into the model for faulty links. In this way, both hard and soft errors in PEs, routers, and links are addressed in the proposed model.

In terms of the failure probability of the links, it could be influenced by many factors such as the adjacent faulty links or the temperature profile of the chip. The detailed model of link being faulty is not the primary concern of this work. The main focus is to make sure that the proposed model is applicable to any type of link failure model. The most sophisticated scenario for any type of model is when the failure probability of each link all differs from each other. Therefore, the failure probability of each link is designed to be able to be assigned separately to satisfy the requirement from different link failure models. When $n(n \leq N)$ out of N links are faulty, there are $M = \binom{N}{n}$ different faulty scenarios due to the different positions of the faulty links. As the metrics (i.e. reliability, energy, performance) of the same mapping pattern varies greatly with respect to different fault scenarios, in this work, all the M conditions are accounted for when optimizing reliability, energy and performance.

B. Reliability, energy and performance co-optimization model

In CoREP, optimization of performance includes both latency and throughput. Latency is evaluated quantitatively to choose the optimal mapping based on three criteria: (1) the mapping with shortest communication path; (2) the mapping with fewest faulty links in the communication path; (3) the mapping with the least congestion. For the third criterion, it is also a major constraint for throughput under the same circumstances. Therefore, throughput is considered qualitatively to choose the mapping with the least congestion [24] which is included in the quantitative analysis of latency.

Energy latency product (ELP) has been proposed to show the energy efficiency at a single injection rate [15]. In this paper, ELP is used to evaluate energy and latency simultaneously. In different application mapping scenarios, the requirement for reliability and ELP varies. For example, the requirement of mobile device is low energy, while the requirement of space systems is high reliability. Therefore, an efficient cost model to be able to distinguish the importance of reliability

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2015.2422843, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems

and ELP is desired when mapping applications onto different systems. Therefore, an efficient cost model must distinguish the importance of reliability and ELP. However, in most cases, reliability enhancement is often obtained at a cost of ELP. Because of this, a weight parameter $\alpha \in [0, 1]$ is introduced to differentiate the priorities of reliability and ELP. In an NoC-based reconfigurable system, reliability is required to be as high as possible, while ELP is required to be as low as possible. Thus, in this study, reliability is measured by reliability cost, which is preferred to be as low as possible. In this way, the overall cost of a mapping pattern can be expressed as

$$Cost = \alpha NR + (1 - \alpha)NELP, \tag{1}$$

where NR and $NELP$ are the normalized reliability cost and normalized ELP respectively. The measurement of reliability cost and energy latency product is discussed in the following discussions.

Reliability cost: Reliability in this study is evaluated by reliability cost: the higher the reliability cost is, the lower the reliability is. The reliability cost of a source-destination pair is defined by a binary indicator of whether there is an available path from source to destination. For example, the required transportation is from R1 to R3 as shown in Fig. 1. If the links numbered 2, 3 and 5 are faulty in Fig. 1(a), there is no available communication path through which the data can be transported from R1 to R3 successfully. Therefore, the reliability cost is defined to be 1. In Fig. 1(b), the links numbered 2, 5 and 18 have errors, but packets can still be transported from R1 to R3 through the path R1$->$R8$->$R3. In this case, the reliability cost is then defined to be 0.

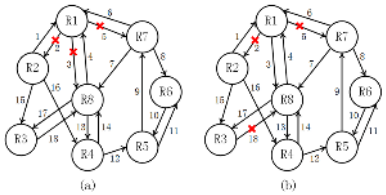


Fig. 1. Two fault patterns when three links are defective.

Under such definition, the reliability cost for the ith condition of all M possibilities when n links are faulty is given by

$$R_{i,n} = \sum\nolimits_{SD} R_i^{SD} F_i^{SD}, \tag{2}$$

where R_i^{SD} is the reliability cost of the source (S)-destination (D) pair under the ith fault condition, and F_i^{SD} indicates whether there is a communication between S and D, as defined by

$$F_i^{SD} = \begin{cases} 1, & a_{SD} \in A \\ 0, & a_{SD} \notin A \end{cases}. \tag{3}$$

Since the reliability cost varies when the fault condition changes, it is important to consider all fault conditions. Let P_i be the probability that the ith fault condition of n faulty links occurs, given by

$$P_i = \prod\nolimits_{j=0}^{n} p_j \times \prod\nolimits_{j=0}^{N-n} (1 - p_j), \tag{4}$$

where p_j is the failure probability of each link and i defines the specific positions of the n faulty links. Hence, the overall reliability cost of a mapping pattern is given by

$$R = \sum\nolimits_{n=0}^{N} \sum\nolimits_{i=1}^{M} R_{i,n} P_i. \tag{5}$$

Then the reliability cost is normalized by a normalization factor N_R, which is obtained by (5) when considering the network that can tolerate the maximum number of faulty links. The normalized reliability cost is defined as

$$NR = R/N_R. \tag{6}$$

The energy latency product (ELP): ELP is obtained by measuring the energy and latency respectively. When referred to energy, static energy and dynamic energy should both be considered. The static energy is mainly influenced by process technology, temperature and supply voltage. For different mapping patterns on a specific chip, only supply voltage and temperature tend to vary within a small range while the process technology is exactly the same. However, as can be seen from the experimental results in [25], the variation of the average temperature for different mappings is about 1 2°C. The maximum relative variance of leakage current of [25] is simulated to be 9.44% using Hspice. According to the static energy model in ORION [7], static energy is proportional to the leakage current. This means that 9.44% variation in static energy will lead to 2.83∼5.66% variation in total energy since the static energy is about 30∼60% of the total energy for 65nm technology [26]. The computation energy, as a part of the dynamic energy, is consumed by PEs for computing the tasks in the applications. It remains unchanged as the overall tasks are the same for different mapping patterns. Unlike these two types of energy, communication energy consumption varies dramatically when the mapping pattern is changed. In addition to that, the communication energy accounts for 28% of the total energy in a router [7], and it can be even more than 40% for most multimedia applications [20]. Therefore, only the communication energy is included in ELP in this work and in the following discussion, energy is referred to communication energy. The bit energy metric introduced in [27] is used to estimate the communication energy of the network. E_{Rbit} and E_{Lbit} indicate the energy consumed by transmitting a bit of data through a router and a link, respectively. Using these two measures, the energy consumed by transporting V_{SD} data from source to destination is given by

$$E_{i,n} = \sum\nolimits_{SD} V_{SD}[E_{Lbit} d_i^{SD} + E_{Rbit}(d_i^{SD} + 1)]F_i^{SD}, \tag{7}$$

where d_i^{SD} is the number of links on the communication path from S to D under the ith fault pattern, and F_i^{SD} is defined in (3).

Various failure conditions of n faulty links can incur different amount of communication energy, and therefore, all the fault scenarios are taken into account when the overall communication energy for a specific mapping is calculated as shown in (8).

$$E = \sum\nolimits_{n=0}^{N} \sum\nolimits_{i=1}^{M} E_{i,n} P_i, \tag{8}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2015.2422843, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems

where P_i is defined in (4), indicating the probability that the ith faulty condition of n faulty links occurs.

In this paper, wormhole is used as the switching technology for the network. In this case, the flit latency of both body and tail flits is the same as that of head flit. For simplicity, in this paper, the flit latency is defined as the time interval between the points that the head flit is established in the source and that it is received by the destination. It includes three parts: (1) the raw communication time by passing the head flit from the source to the destination when there are no faulty links and congestion on its communication path; (2) the waiting time caused by faulty links; and (3) the waiting time caused by congestion. The raw communication time is calculated as the time interval that the head flit is transported from the source to the destination, as expressed in (9).

$$LC_{i,n} = \sum\nolimits_{SD}[t_w d_i^{SD} + t_r(d_i^{SD}+1)]F_i^{SD}. \quad (9)$$

where t_r and t_w represent the time consumption of transporting a flit through a router and a link respectively, and F_i^{SD} is defined in (3). Whenever the head flit meets a faulty link, it is assumed that the head flit will be transported again in the next cycle. The transportation of the head flit will be tried for the every following cycle until the fault in the link is removed. Therefore, the waiting time caused by a faulty link j is estimated by the average wait time as expressed in (10).

$$LF_j = \lim_{T\to\infty}(p_j + 2p_j^2 + 3p_j^3 + \cdots + Tp_j^T) = \frac{p_j}{(1-p_j)^2}, \quad (10)$$

where p_j is the failure probability of link j and T means the cycles the head flit must wait. Congestion is addressed by a first-in-first-out (FIFO) queue, and each router is regarded as a server in the queue. Under the deterministic routing algorithms, the packet is transported to the definite router when the source and destination are determined. In this case, there is only one server in each queue. However, as to adaptive routing algorithms, the packet can choose which router to transport according to the current state of the network. This means that multiple servers are ready to serve this packet. Accordingly, the waiting time due to congestion is estimated by the G/G/m-FIFO priority queue, as the interarrival time and service time are both regarded as independent general distributions. Using the Allen-Cunneen formula [28], the waiting time of the uth input port to the vth output port of router K can be calculated by (11) (13).

$$WT_{u\to v}^K = \frac{\overline{WT_0^K}}{(1-\sum_{x=u}^{U}\rho_{x\to v}^K)(1-\sum_{x=u+1}^{U}\rho_{x\to v}^K)}. \quad (11)$$

$$\overline{WT_0^K} = \frac{P_m}{2m\rho} \times \frac{C_{A_{u\to v}^K}^2 + C_{S_v^K}^2}{\mu_v^K} \times \rho_v^K. \quad (12)$$

$$P_m = \begin{cases} \frac{\rho^m+\rho}{2}, & \rho > 0.7 \\ \rho^{\frac{m+1}{2}}, & \rho < 0.7 \end{cases}. \quad (13)$$

In (12), $C_{A_{u\to v}^K}^2$ is the coefficient of variation of the arrival process to the router K. As the arrival process to each router is assumed to be the same as that to the network, $C_{A_{u\to v}^K}^2$ is equal to the coefficient of variation of the arrival process to the network (C_A^2), and it is determined by the $APCG$. Similarly, $C_{S_v^K}^2$ is the coefficient of variation of the service process on the router K. As shown in Fig. 2(a), the service time at output port i of R4 consists of three parts: 1) raw service time to pass through R5 without congestion; 2) time spent on waiting the arbitration from the input j to the output k; 3) time spent on waiting the output port k to be free, that is, the service time at output k of R5. Since each output port of R5 has a great impact on the service time at output port i of R4, an interdependency tree is established to deal with that. When establishing the interdependency tree, the router connecting to k is added on the tree if there are communications on the output port k of R5; otherwise, it is not added as shown in Fig. 2(b). This establishment will continue until the router only communicates with its corresponding PE. After the interdependency is established, the service time of the leaf nodes are firstly calculated, and then the service time of their parent nodes are computed, as shown in (14). In (14), $\overline{S_v^K}$ represents the average service time at output port v of router K and $\overline{(S_v^K)^2}$ represents the average second moment of service process.

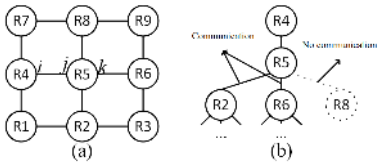


Fig. 2. (a) Example of one topology; (b) Interdependency tree corresponding to the topology.

$$\overline{S_v^K} = \sum\nolimits_{x=1}^{Q} \frac{\lambda_{u\to x}^K}{\lambda_x^K} \times (t_r + t_w + WT_{u\to x}^{K+1} + \overline{S_x^{K+1}} - t_b),$$

$$\overline{(S_v^K)^2} = \sum\nolimits_{x=1}^{Q} \frac{\lambda_{u\to x}^K}{\lambda_x^K} \times (t_r + t_w + WT_{u\to x}^{K+1} + \overline{S_x^{K+1}} - t_b)^2,$$

$$C_{S_v^K}^2 = \frac{\overline{(S_v^K)^2}}{\overline{(S_v^K)}^2} - 1. \quad (14)$$

The parameters used in (11) $\sim$ (14), which are determined by the APCG and the structure of routers, are defined in Table I. Then the latency for the ith fault condition of n links being faulty is calculated by

$$L_{i,n} = LC_{i,n} + \sum\nolimits_{SD}[\sum\nolimits_{K=1}^{d_i^{SD}} WT_{U(K)\to V(K)}^{R(K)} + \sum\nolimits_{j=1}^{d_i^{SD}+1} LF_{L(j)}]F_i^{SD}, \quad (15)$$

where $R(K)$ is the function to obtain the index of the Kth router on the communication path of S and D, $U(K)$ and $V(K)$ are the functions to obtain the index of the routers input port and output port, and $L(j)$ is the function to obtain the number of the jth link. When all the faulty conditions are taken into account, the total latency is then calculated by

$$L = \sum\nolimits_{n=0}^{N}\sum\nolimits_{i=1}^{M} L_{i,n}P_i. \quad (16)$$

Energy latency product is then obtained and normalized by

$$NELP = E \times L/N_{ELP}, \quad (17)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2015.2422843, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems

TABLE I
PARAMETERS USED IN CALCULATING LATENCY

ρ_v^K, ρ	The fraction of time that the vth output port of router K is occupied by its xth input port ($\rho_v^K = \sum_{x=1}^{P} \lambda_{x\to v}^K / \mu_v^K$, $\rho = \sum_{v=1}^{Q} \rho_v^K$).
m	The number of the candidate routers.
C_A^2	Variation for Interarrival time of packets.
$\lambda_{x\to v}^K$	Average flit rate(flit/cycle).
μ_v^K	Average service rate (cycle/flit).
P	The amount of input ports of a router.
Q	The amount of output ports of a router.
t_b	The time consumed by passing the buffer.

where the normalization factor N_{ELP} is computed by multiplying (8) by (16) when assuming the worst case of passing the most number of nodes in the communication path.

The CoREP is applicable to various NoC topologies and routing algorithms, because of the following reasons. Firstly, the proposed approach only counts the number of faulty links in all communication paths, and then chooses the mapping with the least number of faulty links; this process is applicable to a broad category of NoC topologies. However, other literature models reliability by assuming the bounding box of a source-destination pair to be closer to a square [18], which is limited to a mesh topology. Secondly, the energy is measured dynamically in CoREP, so the evaluation can be done as soon as the communication path changes. This process is independent of the NoC architectures and is different from the previous models. The previous model in [18] requires the communication path in advance, which is limited to deterministic routing algorithms; while the model in [19] requires to know the Manhattan distance of the source-destination pair [19], which is limited to mesh topology. Thirdly, the G/G/m-FIFO priority queuing model used in evaluating latency ensures that this model is applicable to diverse NoC topologies and routing algorithms. In conclusion, the proposed approach is flexible in terms of NoC topologies and routing algorithms with the awareness of reliability, energy and latency, which is an improvement over previous approaches.

III. CoREP Oriented Mapping Approach

A. Problem definition

Using CoREP, the problem of a reliability-, energy-, and latency-aware mapping is defined as follows.

Given an application characteristic graph APCG and an NoC of routers and PEs with any topology and routing algorithm;
Find a mapping function map() that maps an IP core $c_i \in C$ in the $APCG$ to a PE in the NoC;
Such that the following conditions are satisfied:
Min: $Cost = \alpha NR + (1-\alpha)NELP$;
S.t. $map(c_i) \neq map(c_j), \forall c_i \neq c_j \in C$.

B. Priority and ratio oriented branch and bound mapping

Branch and bound mapping method: Branch and bound (BB) method is a widely used approach for Non-deterministic Polynomial problems (NP-problems) [29]. In BB, the minimum of a cost function is calculated by establishing a search tree. An example of the search tree is shown in Fig. 3, which shows the mapping flow of an APCG with 3IPs onto an NoC. For NoC mapping, each node in the search tree indicates a candidate mapping with the mapping order of IPs. The numbers in the nodes indicate the indices of the IPs while the positions of these numbers correspond to the indices of the nodes in the NoC. The blank space in the nodes means that no IP has been mapped onto this node yet. In this way, the root node with three blank spaces represents the mapping that no IP is mapped which is the start point of mapping flow. As the mapping proceeds, IPs are mapped onto the first node of the NoC and then other nodes in turn. These partial mappings that only some of IPs are mapped are represented by the internal nodes. For example, the internal node "31_" means that IP3 and IP1 are mapped onto the first and second node of the NoC, respectively. The mapping flow is terminated when the leaf nodes are generated with all IPs mapped on the NoC. When the search tree is established, lower bound cost (LBC) and upper bound cost (UBC) of the internal nodes are utilized to decide whether the internal nodes should be created or not. If LBC of an internal node is larger than its minimal UBC, this internal node and all its corresponding child nodes are discarded. In this way, the mappings, which are less likely to be the optimal solution, are discarded in the early stages to reduce computational complexity.

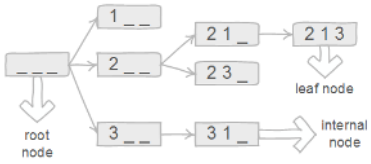


Fig. 3. An example of the search tree.

Priority and ratio oriented branch and bound mapping: Following the branch and bound mapping approach, a priority and ratio oriented branch and bound mapping approach (PRBB) is proposed to map an application onto an NoC efficiently. Two techniques, branch node priority recognition technique and partial cost ratio utilization, are adopted to raise the efficiency of finding the best mapping pattern.

1) Branch node priority recognition technique: According to the search tree defined to represent the candidate mappings, the internal nodes that are closer to the root node lead to more computational overhead. Therefore, in PRBB, the nodes with a shorter distance to the root node are assigned a higher priority. During the mapping procedure, the priorities of the internal nodes are first recognized; then the nodes with higher priorities are addressed in order of priority. If these nodes are not likely to be the optimal mappings, they are discarded in early stages and the search efficiency is improved.

2) Partial cost ratio utilization technique: Partial cost ratio between two adjacent costs is defined in (18).

$$ratio_{n+1,n} = \frac{Cost^{n+1}}{Cost^n}, \tag{18}$$

where $Cost_n = \sum_{i=1}^{\binom{N}{n}} [\frac{\alpha R_{i,n}}{N_R} + \frac{(1-\alpha)E_{i,n}L_{i,n}}{N_{ELP}}]P_i$, denoting the cost with n faulty links, which is defined as the nth partial cost.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2015.2422843, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems

As discussed previously, CoREP is applicable to links with multiple failure probabilities. To demonstrate the non-unified failure probabilities, two values, p_h and p_l are used in this paper as an example for simplicity. The difference is based on the communication volumes passing through the link, because larger volumes lead to higher energy consumptions. Moreover, high energy consumption is likely to result in thermal hotspots and high temperature is more likely to cause errors in links. Therefore, when the link is in the region of predefined high communication volumes, p_j is considered to be p_h; otherwise it is p_l. This is a simple example to show the variance of failure probabilities and the proposed model can also be integrated with more sophisticated thermal models. Then P_i is calculated by substituting p_j with p_h and p_l in (4).

When n changes to $n+1$, the variation of $R_{i,n}$ is limited to a small set of paths. Moreover, $E_{i,n}$ changes very little as the number of links on the communication path only changes slightly. The waiting time caused by the faulty links is counted as the average value, therefore the change of $L_{i,n}$ is small as well. In this way, (18) can be simplified to

$$ratio_{n+1,n} < \frac{N-n}{n+1} \times \frac{1}{4(1-p_l)^2}. \quad (19)$$

As the failure probability of one link is typically smaller than 0.5, $ratio_{n+1,n}$ decreases rapidly when n becomes large; therefore, $ratio_{n+1,n}$ is small enough to be ignored with a large n. Moreover, $Cost^n$ can be calculated by

$$Cost^n = Cost^0 \prod_{k=1}^{n} ratio_{k,k-1}. \quad (20)$$

Hence, the overall cost, which is the sum of all the partial cost, can be simplified by adding up the first several ones. This technique can reduce the computational overhead.

In PRBB, the condition to delete a non-optimal candidate mapping is defined by (21) to ensure the accuracy of the optimal solution.

$$LBC > min\{UBC\} \times (1 + ratio_{1,0}), \quad (21)$$

In (21), LBC means the lower bound cost, which is calculated by adding up three parts: 1) the cost among the mapped nodes, 2) the cost among the unmapped nodes, and 3) the cost among the mapped and unmapped nodes. UBC means the upper bound cost, which is the cost of a leaf node created by a temporary greedy mapping of the remaining nodes. If the condition defined in (21) is met, the node and all its child nodes are discarded, otherwise the node is saved for further comparison. By using this deleting condition, the internal nodes with a similar overall cost will be saved for further comparison. This ensures that the accuracy of the optimal result will not be sacrificed when speeding up the search procedure.

Work flow of application mapping onto NoC-based reconfigurable computing system using CoREP and PRBB: Since the proposed approach is designed for NoC-based reconfigurable computing system, the work flow of the overall application mapping using CoREP and PRBB is shown in Fig. 4. For a given application and a certain NoC, the optimal mapping pattern is firstly given by PRBB based on the overall cost given by CoREP. During the application running on the NoC, the topology and/or routing algorithm reconfiguration of the NoC is required in case of special events such as the occurrence of permanent faults or application requirement [22][30]. After the instantaneous topology reconfiguration, a remapping procedure with PRBB is implemented and another optimal mapping pattern corresponding to the new topology and/or routing algorithm is figured out at run-time.

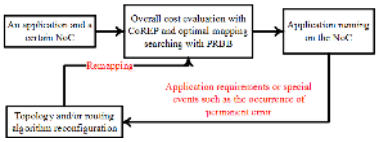


Fig. 4. Work flow of application mapping onto NoC-based reconfigurable computing system using CoREP and PRBB.

Computation Complexity: After discussing the two techniques used to reduce the computational overhead in PRBB, the computational overhead is estimated and compared to BB. It is difficult to figure out the accurate number of internal nodes and when the nodes are discarded, we make an assumption that k nodes remain at each branch in PRBB. Compared to BB, PRBB focuses on discarding more nodes closer to the root node. This means that BB suffer from larger computation overhead and therefore BB is assumed to have one more node left on average. As each loop contains almost the same basic operations, the number of loops is regarded as the time complexity of an algorithm, which is adopted to quantify the computational complexity of the algorithm. Subsequently, the results for PRBB and BB are shown in (22) and (23) respectively,

$$CC_{PRBB} = o(m^3) \times (\frac{(k-1)^{m+1} - k + 1}{(k-2)^2} - \frac{m}{k-2}), \quad (22)$$

$$CC_{BB} = o(m^3) \times (\frac{k^{m+1} - k}{(k-1)^2} - \frac{m}{k-1}), \quad (23)$$

where m means the nodes of an NoC. The ratio of the computational overhead (BB/PRBB) is plotted with Matlab and shown in Fig. 5. It can be seen that the ratio of the computational complexity is more than 1 in all cases, which indicates that PRBB can find the best mapping with a shorter time compared to BB. It can also be seen from Fig. 4 that the ratio increases when the scale of NoC increases, which indicates that PRBB is preferable for NoCs with large scale.

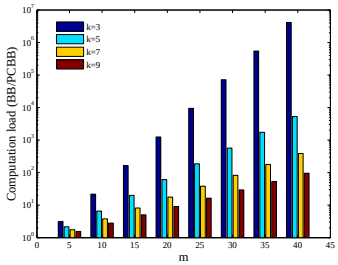


Fig. 5. Computation complexity reduction of PRBB against BB.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2015.2422843, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems

IV. Experimental Results

In this section, experiments are performed to verify the flexibility and accuracy of CoREP and the efficiency of PRBB. Firstly, experiments are done to address the mapping issues on four different combinations of topology and routing algorithm. It demonstrates that CoREP can actually work on these different topologies and routing algorithms. Secondly, comparisons with one state-of-the-art approach [18] and a classical algorithm under a specific topology and routing algorithm is performed.

The best mapping, as well as the run time, is reported by PRBB in C++. Using the best mapping pattern, experiments are done on a cycle accurate simulator implemented by SoCDesigner [31]. Each node of the NoC in this simulator contains a router and a PE, and the experiment environment in a router is shown in Table II. As both the proposed cost model and mapping approach are independent of the switch technology, arbitration policy and the use of virtual channels, the parameters are chosen to be the same as the approach in [18]. In the simulation, errors are injected randomly into the NoC with two probabilities, p_l and p_h, depending on the position of the links. Reliability is then estimated as the probability of transporting a flit accurately from its source to its destination [6]. As a probabilistic measurement, the unit of reliability is set to be 1 in this simulation. The communication energy for a specific mapping pattern is calculated by the following two steps: (1) counting the numbers of routers and links in each communication path; (2) multiplying the numbers by E_{Rbit} and E_{Lbit} respectively. The values of E_{Rbit} and E_{Lbit} are from an open source simulator [32], which is gained with Synopsys Power Complier using the model in [27], as shown in Table II. The total energy is then divided by the number of transported flits to obtain an average energy consumed by passing one flit [5]. One packet in this simulation is assumed to have eight flits and the flit latency is calculated as the time interval between the head flit being established in the source and being received by the destination. Finally, throughput is evaluated as the average flits each node delivers during the simulation.

TABLE II
Experimental Environment For Routers

Switch technology	Wormhole
Arbitration policy	Roll-turn
Virtual channel	Use virtual channels
E_{Rbit}	4.171 nJ/bit [32]
E_{Lbit}	0.449 nJ/bit [32]

A. Verification for Flexibility

Four different combinations of NoC topology and routing algorithm are used in this subsection to verify the flexibility of our proposed cost model. The choice of the topologies and routing algorithms are based on a survey of 60 literatures including 66 topologies and 67 routing algorithms [33]. It shows that about 56.1% of the networks are mesh/torus, 12.1% are custom and 7.6% are rings, while 62.7% of the networks use deterministic routing algorithms and the remaining 37.3% use adaptive routings. Therefore, the topologies chosen for experiments are torus, Spidergon, de Bruijn Graph and mesh, and the corresponding routing algorithms are oddeven, crossfirst, deflection and full-adaptive, as shown in Table III. Eight benchmarks, representing eight different real applications, are used in the experiment. The information for the eight benchmarks is shown in Table IV. The first four are widely used benchmarks, which are generated from real applications [34]. The other four applications with substantial communication volumes are chosen to satisfy the increasing complexity requirement of recent NoC-based reconfigurable computing systems. H264[35] and HEVC[36] are two complex and state-of-the-art video coding standards, while freqmine and swaption are generated from the Princeton Application Repository for Shared-Memory Computers [37]. The number of IPs for the former four applications is chosen to be the same as that in [34] while for the latter four it is based on a common criterion of balancing the communication volume between IPs. The network size is then chosen to satisfy both the minimum number requirement based on the application and the specific requirement from the topology of the NoC.

As the aim of the current experiment is to show the flexibility of CoREP, it is not required to lay different emphasis on energy latency product or reliability; therefore α is set to 0.5 in this experiment. When the failure probability of a link is larger than 0.5, the NoC is impossible to work effectively; therefore, the failure probabilities are chosen to be no larger than 0.5 (i.e. p_l = 0.5, 0.1, 0.01, 0.001, 0.0001 and p_h = 0.5, 0.5, 0.1, 0.01, 0.001). The similar literatures lacks the study on simultaneously addressing reliability, energy and latency on various NoC topologies and routing algorithms. Moreover, the run time of exhaustive search to obtain the global optimal mapping is too long to be acceptable. Therefore, PRBB is compared to a classical mapping algorithm using simulated annealing (SA) [38], which is a probabilistic method for

TABLE III
Topology and Routing Algorithms Combinations Used for the Demonstration of Flexibility

No.	Topology		Routing Algorithm	
	Name	Category	Name	Category
1	Torus	Mesh/Torus	Odd-even	Adaptive
2	Spidergon	Ring	CrossFirst	Deterministic
3	deBruijnGraph	Custom	Deflection	Deterministic
4	Mesh	Mesh/Torus	Full-adaptive	Adaptive

TABLE IV
Benchmarks Used in the Simulation

Benchmark	Number of IPs	Application
VOPD	16	Video object plane decoder
MWD	12	Multi-window display
PIP	8	Picture in picture
DVOPD	32	Dual video object plane decoder
H.264	14	H.264 decoder
HEVC	16	High Efficiency Video Coding decoder
freqmine	12	Data mining application
swaption	15	Computes portfolio prices using Monte-Carlo simulation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2015.2422843, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems

finding the global minimum of a cost function that may possess several local minima. SA is just a mapping algorithm, which is flexible, but not reliability, energy or latency aware. Moreover, a different cost model incurs a different computational complexity during a mapping, therefore SA is considered to use the same cost model as PRBB for a fair comparison.

All the eight benchmarks in Table IV are mapped onto the four combinations of NoCs in Table III with PRBB and SA. For each benchmark mapped onto an NoC, five combinations of link failure probability are utilized (p_l = 0.5, 0.1, 0.01, 0.001, 0.0001 and p_h = 0.5, 0.5, 0.1, 0.01, 0.001, respectively). These experiments clearly show the flexibility of CoREP. Moreover, the average results for all eight benchmarks and four NoCs with respect to the link failure probability is shown in Fig. 6. The average reliability enhancement is shown in Fig.6(a). When pl is small enough, any mapping pattern is highly reliable. Consequently, the improvement in reliability becomes smaller as pl reduces. Fig. 6(b) shows the average run time ratio (SA/PRBB) with respect to the failure probability of a link. It can be seen that the processing time of SA is at least 500x of PRBB. This advantage is attributed to the two techniques used in PRBB. It can also be seen that when the failure probability decreases the ratio becomes large. The reason for that is a smaller failure probability of a link means fewer faulty links in the network and PRBB will spend less time dealing with the faulty links during the mapping. On the other hand, SA computes all the candidate mapping patterns, so the run time to find the best mapping pattern is independent of the failure probability of a link. The average energy reduction is depicted in Fig. 6(c) and for all cases the best mappings found by PRBB outperform those found by SA in terms of energy consumption. Although the energy consumption will increase when p_l increases, as a larger link failure probability means more energy is consumed by addressing the faulty links, the margin of energy reduction is independent of the failure probability because of the utilization of CoREP. Latency is closely related to throughput, which is evaluated in the next section.

All these results show that the cost model, CoREP, is independent of, and therefore applicable to diverse types of NoC topologies and routing algorithms. This ensures the flexibility of the proposed mapping approach. Moreover, PRBB can find a much better mapping pattern with a considerable reduction in computation time compared to SA, for the same cost models.

B. Comparisons under a specific topology and routing algorithm

In this subsection, experiments are performed for further comparisons on a specific topology and routing algorithm. PRBB is firstly compared to a state-of-the-art approach with a different cost model and mapping approach, with respect to reliability, energy, latency, throughput and computational complexity. As discussed before, the methods in [18], [19] have considered two metrics, however they are both limited to a specific topology and routing algorithm. Moreover, the approach in [19] considers the different condition from that in this work, so the branch and bound (BB) method in [18] is chosen as the baseline for comparison. To make a fair comparison, the topology and routing algorithm used in the simulator are the same as those used in [18]. The characteristics of the eight benchmarks and the corresponding NoC size are shown in Table V. The first four benchmarks are chosen to be the same as those in [18] and the latter four are chosen based on the same reason as explained for Table IV.

TABLE V
CHARACTERISTICS OF BENCHMARKS AND CORRESPONDING NOC SIZE

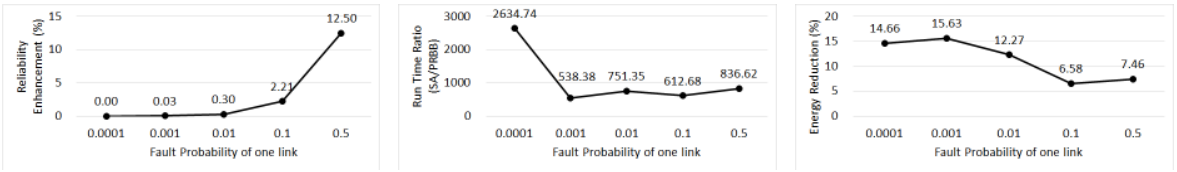
Benchmark	Number of IPs	Min/Max communication	NoC Size
mpeg4	9	1/942	9
telecom	16	11/71	16
ami25	25	1/4	25
ami49	49	1/14	49
H.264	14	3280/124417508	16
HEVC	16	697/1087166	16
freqmine	12	12/6174	16
swaption	15	145/747726417	16

Extensive experiments are carried out on the simulator implemented by SoCDesigner to evaluate the best mapping patterns obtained by PRBB and BB. A summary of 50,000 fault injection experimental results, including the maximum, minimum and average values for two different weights, are shown in Table VI. In addition, all the experiments are also run using SA, and the comparisons are also shown in Table VI. It can be seen that PRBB outperforms both BB and SA under every aspect. The discussion when $\alpha = 0.2$ is given subsequently in detail.

Reliability enhancement: The comparisons of the reliability with respect to the link failure probability are depicted in Fig. 8. Compared to BB, 6 out of the 8 benchmarks show noticeable improvement while the improvement for ami25 and ami49 is small. This is because the communication volume of each source-destination pair is smaller than others (in Table V), and fewer communication paths are occupied. In this case, alternative communication paths are available when errors occur in the original path, so it is difficult to obtain a significant improvement in reliability. For the other 6 benchmarks, the reliability enhancement when $p_l \leq 0.025$ is quite small as any mapping pattern is highly reliable with a low failure probability. However, the reliability improvement when $p_l \geq 0.04$ is remarkable, which confirms the advantage of PRBB. To summarize, the reliability of the best mappings found by PRBB is on average 10% higher than that of BB. This is because when reliability cost is considered in CoREP, all the faulty conditions are taken into account, so it is more accurate. However, BB only considers the bounding box of the source-destination pair to be close to a square, which is independent of the fault conditions.

In certain circumstance, the reliability of the network has the highest priority but the fault-rate is quite high, such as space environment. The proposed cost model CoREP can place the emphasis on reliability by changing the weight parameter α as required. In the meanwhile, energy and performance are sacrificed inevitably as a cost. To quantify the outcome and the sacrifice of the emphasizing on reliability, experiments are divided into two parts. Firstly, the baseline for comparison is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2015.2422843, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems



(a) reliability enhancement (b) run time ratio (c) energy reduction

Fig. 6. (a) Average reliability enhancement, (b) run time ratio (SA/PRBB), and (c) energy reduction, with respect to the failure probability of a link(p_l).

TABLE VI
SUMMARY OF COMPARISONS AMONG PRBB, BB AND SA

	$\alpha = 0.2$						$\alpha = 0.6$					
	Against BB			Against SA			Against BB			Against SA		
	Max	Min	Avg.	Max	Min	Avg.	Max	Min	Avg.	Max	Min	Avg.
Energy Reduction	40%	4%	24%	59%	-13%	28%	47%	-1%	25%	57%	-13%	25%
Reliability Enhancement	106%	0.01%	10%	208%	-4%	12%	107%	-0.01%	10%	233%	-2%	16%
Latency Reduction	49%	4%	17%	40%	0.8%	20%	46%	5%	16%	42%	-0.2%	20%
Throughput Improvement	22%	5%	9%	22%	4%	9%	22%	6%	10%	22%	4%	10%
Computation Time Reduction	20x	1x	3x	4477x	111x	1041x	27x	1x	4x	5584x	137x	972x

made by finding the optimal mapping pattern when the fault rate is very low which is set to 0.0001. Then the energy, performance and reliability of this mapping pattern are evaluated for six fault-rates ranging from 0.0001 to 0.5 respectively. The reduction in reliability when fault-rate increases is shown by the blue curves in Fig. 7. In the second part of the experiments, for each fault-rate the mapping with the greatest reliability is found with the proposed approach. The same evaluation is made which shows the increment in energy and latency (green curve in Fig. 7) as well as the reduction in reliability (red curve in Fig. 7). For both parts, eight benchmarks are tested and experimental results show that the reliability is increased by 16.5% while the energy latency product increases by 54% on average. In addition, with the proposed approach, the improvement of reliability reaches up to 104% when the fault-rate increases.

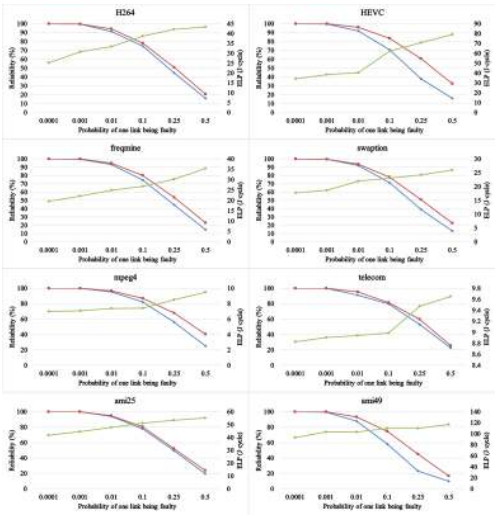


Fig. 7. Reliability and energy latency product (ELP) variation with fault probability increases. Blue: Reliability with ELP unchanged, Red: Reliability with ELP sacrificed; Green: The absolute value of ELP.

Energy reduction: The energy consumption is compared

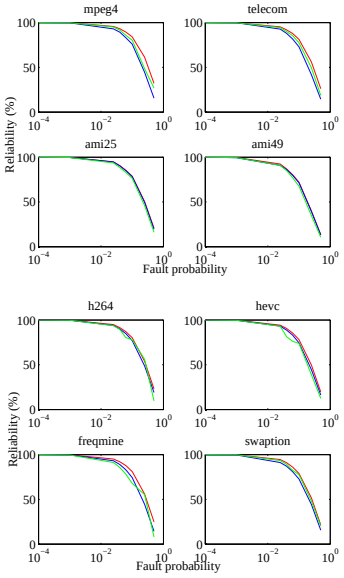


Fig. 8. Reliability comparison of the best patterns found by PRBB (Red), BB (Blue) and SA (Green) for different benchmarks.

in Fig. 9 for the best mappings on the NoC. For all of the eight benchmarks, PRBB consumes less energy than BB. On average, the energy reduction obtained by PRBB is about 24% compared to BB. This is mainly because that in CoREP, energy is calculated by considering the effects of all fault patterns. In addition to that, the contribution of communication energy reduction to the total energy is further discussed. The range of improvement in communication energy is identified by comparing the optimal mapping found by PRBB with a random mapping which represents the worst case for communication energy. Then this range is converted to the changing range of the total energy. As shown in Table VII, the range of eight benchmarks is compared based on the assumption that the communication energy takes up 28% of total energy consumption [7]. It shows the contribution of communication energy is between 14.3% $\sim$ 31.0%, which is a smaller than the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2015.2422843, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems

actual ratio because the random case might not be the worst case. Compared with the contribution of static energy (2.83 $\sim$ 5.66%), the communication energy is of significant importance to the overall energy consumption which also confirms the theoretical analysis in the energy model.

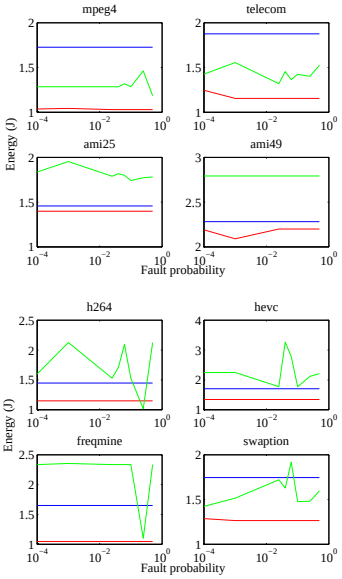


Fig. 9. Communication energy comparison of the best patterns found by PRBB (Red), BB (Blue) and SA (Green) for different benchmarks.

Performance: Performance, in terms of latency and throughput, is also evaluated for a comprehensive comparison. As latency is closely related to throughput, it is compared with respect to throughput using the best mappings found by PRBB and BB when $p_l = 0.01$ and $p_h = 0.1$. As shown in Fig. 10, the latency for each benchmark remains 20 $cycles/flit$ approximately when the throughput is small. However, as the throughput increases, the network comes into saturation, resulting in a latency wall. On average, PRBB outperforms BB with about 17% reduction in latency. The latency reduction is due to the quantitative evaluation of CoREP. For each mapping pattern, the time consumed by passing flits on the communication path and the waiting time caused by faulty links and congestion are both estimated in CoREP. Therefore, the mapping pattern that incurs a large latency is less likely to be reported as the optimal one.

The comparison of throughput with respect to the failure probability is shown in Fig. 11. It is clear that PRBB outperforms BB and gains about 9% improvement in maximum throughput. Although throughput is not modeled quantitatively in CoREP, it is accounted for in a qualitative evaluation of the latency, as discussed in section II. Therefore, the best mapping found by PRBB, which incurs a low latency overhead, also achieves a high throughput.

The experiment about the actual throughput against injected throughput is also done to supplement the throughput analysis as shown in Fig. 12. For different faulty probabilities, the actual throughput of the NoC is the same when the injected throughput is small. As expected, the actual throughput for all faulty probabilities increases as the injected rate increases

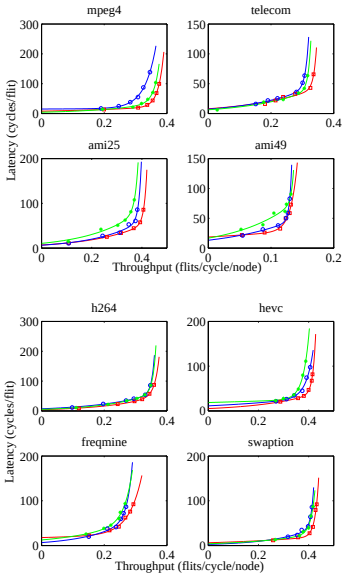


Fig. 10. Latency comparison of the best patterns found by PRBB (Red), BB (Blue) and SA (Green) when $p_l = 0.01$ and $p_h = 0.1$ for different benchmarks. (Markers: simulated data, Line: fitting curve).

until it reaches saturation which is the maximum throughput of the NoC [24]. According Fig. 12, the results in Fig. 11 all fall in the range of saturation with an injected rate of 0.5 flit/cycle/node. Corresponding to the results in Fig. 11, the actual throughput at saturation decreases when the faulty probability increases.

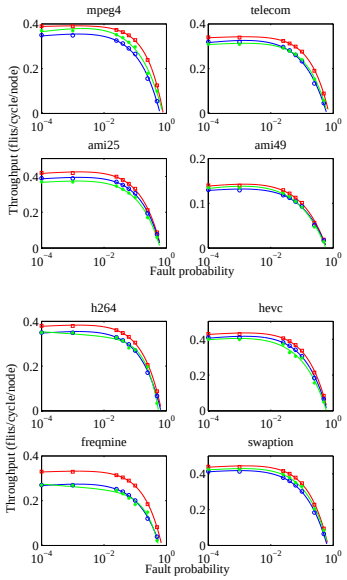


Fig. 11. Throughput comparison of the best patterns found by PRBB (Red), BB (Blue) and SA (Green) for different benchmarks with the injected rate of 0.5 flit/cycle/node. (Markers: simulated data, Line: fitting curve).

Run time reduction: The run time spent in finding the best mapping pattern is utilized to approximate the computational complexity of each approach. For a fair comparison, the programs of PRBB and BB are run on the same platform, as described in Table VIII. The comparisons are shown in Fig. 13.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2015.2422843, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems

TABLE VII
CONTRIBUTIONS OF COMMUNICATION ENERGY TO TOTAL ENERGY

	mpeg4	telecom	ami25	ami49	H264	HEVC	freqmine	swaption
Communication energy of optimal mapping (J)	1.03	1.15	1.40	2.19	1.15	1.34	1.05	1.29
Communication energy of random mapping (J)	1.93	1.85	2.66	3.81	1.73	2.20	2.21	2.14
Variation in communication energy (%)	87.1	60.0	90.2	73.8	51.0	64.2	110.6	63.4
Variation in total energy (%)	24.4	16.8	25.3	20.7	14.3	18.0	31.0	18.6

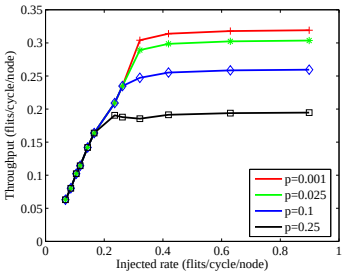


Fig. 12. Actual throughput vs injected rate at different faulty probabilities for benchmark freqmine.

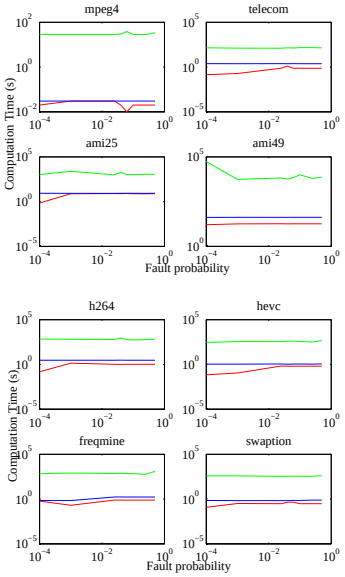


Fig. 13. Comparison of run time to find the best patterns by PRBB (Red), BB (Blue) and SA (Green) for different benchmarks.

As can be seen, for all the eight benchmarks PRBB found the best mapping pattern with a shorter run time than BB. More specifically, the run time spent by BB is approximately 3x of that of PRBB. This speedup is due to the branch node priority recognition and partial cost ratio utilization techniques used in PRBB. The first technique helps to discard as many branch nodes as possible that are closer to the root node to reduce computational complexity. Meanwhile, the second technique estimates the overall cost of a mapping pattern without calculating all the partial costs to control computation overhead. However, BB lays no emphasis on the nodes closer to the root node, which results in a larger computational overhead. Moreover, it calculates all the partial costs to accumulate the overall cost of a mapping pattern.

TABLE VIII
PLATFORM FOR RUNNING MAPPING APPROACHES

CPU	2 × Inter(R) Xeon(R) E5520
Main Frequency	2.27GHz
Memory	16GB
Operating System	Linux

Further computation time comparison is carried out with a method dedicated for run-time optimization [21], and the results are shown in Table IX. In this experiment, the same number of faulty components is chosen as [21]. In Table IX, "F" means the number of faulty cores for [21] while "L" means the number of faulty links for PRBB. Table IX shows that although PRBB consumes more time than LICF [21], PRBB is much faster than MIQP [21]. This could confirm the efficiency of the proposed PRBB because the searching space for PRBB is much larger than LICF and MIQP. In other words, much more candidates need to be searched for PRBB than LICF/MIQP when dealing with the same number of faulty components. Still, the computation time of PRBB and LICF are of the same order of magnitude.

All these results show a remarkable improvement in terms of reliability, energy, latency, throughput and run time when comparing PRBB to BB. At the same time, PRBB also outperforms SA in all these metrics, as shown in Figs. 8 to 13.

TABLE IX
EXECUTION TIME COMPARISON

App	Mesh size	F	Time (sec)		L	Time (sec)
			LICF[21]	MIQP[21]		PRBB
Auto-Indust (9 IPs)	4 × 4	2	0.01	0.2	2	0.04
		4	0.02	2.51	4	0.04
		6	0.04	51.62	6	0.05
		7	0.04	177.72	7	0.06
TGFF-1 (12 IPs)		2	0.01	0.44	2	0.03
		3	0.02	1.34	3	0.05
		4	0.03	4.30	4	0.05

Although PRBB is designed to co-optimize three metrics, it can also be applicable and efficient when it comes to the co-optimization of any two of them. To confirm that, experiments with only energy and performance are carried out by setting α to be zero. PRBB is compared to SA and BB [5] which optimizes energy under performance constraints. Eight benchmarks in Table V are mapped on a 2D mesh NoC with XY routing algorithm. The comparisons are among energy, throughput, latency and run time. Experimental results in Table X show that PRBB outperforms BB and SA in almost all cases. Although SA can sometimes find a superior mapping pattern than PRBB does, it takes too long to find it which is not acceptable in reconfigurable computing systems. Along

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2015.2422843, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems

with previous comparison with other approaches aiming at co-optimization of energy and reliability in Table VI, it can be concluded that PRBB can also work efficiently when only two metrics are considered. This also means that PRBB can be applicable to a wide range of application scenarios, and it has overall advantages in most cases.

TABLE X
COMPARISONS AMONG PRBB, BB AND SA WHEN $\alpha = 0$.

	Against BB			Against SA		
	Max	Min	Avg.	Max	Min	Avg.
Energy Reduction	48%	0%	21%	40%	-4%	11%
Throughput Improvement	37%	0%	14%	34%	-22%	5%
Latency Reduction	19%	8%	13%	22%	6%	12%
Computation Time Reduction	3.3x	0.6x	1.8x	51020x	664x	9687x

C. Discussion about α, NoC scaling and NoC structure

In CoREP, the weight parameter α is used to make the tradeoff between reliability and ELP, therefore, the value of α is of great importance. In the previous simulations, the value of α is chosen to be the same as BB for fair comparisons. In this subsection, an example of benchmark freqmine when $p_l = 0.01, p_h = 0.1$ is taken to show the response to changes in α, and the results are shown in Fig. 14. As the theoretical analysis, experimental results show that the reliability increases (or remains the same) as α increases. However, the ELP is also increasing which is sacrificed during the reliability enhancement. To meet different requirement in the real applications, α should be chosen properly and a tradeoff between reliability and ELP can then be achieved.

The computation complexity is closely related to the scale of NoC, i.e. the number of nodes in the network. The computation complexity of the proposed PRBB with respect to different scale of NoC is investigated in Fig. 15. As a comparison reference, experiments for the exhaustive searching are also carried out. As expected, the computation complexity represented by the run time, of both methods increases as the scale of NoC increases. However, the rate of increase for PRBB is much less than that of exhaustive searching with the benefit of those techniques discussed in Section III.B. This confirms the efficiency of the proposed PRBB even for large scale NoCs.

The discussion of the proposed approach has been based on the commonly used time division multiplexing NoC (TDM-NoC). However, the approach can also be extended to the upcoming spatial division multiplexing NoC (SDM-NoC) which has considerable advantages in energy and performance in some special cases [39]. The fundamental reason for that is the primary difference between TDM-NoC and SDM-NoC is the structure of the NoC routers while the cost model CoREP is based on the system-level. For the evaluation of reliability of a source-destination pair, the reliability cost could also be utilized. Defined as a binary indicator of whether or not there is an available path from source to destination, the reliability cost is independent of the structure of the router. In this way, the calculation of reliability cost for TDM-NoC remains unchanged. As in (7), energy is calculated by summing up the energy consumed by routers and links on the communication path. Since the structure of the routers of SDM-NoC differs from that of TDM-NoC, the energy consumed by each router E_{Rbit} and each link E_{Lbit} needs to be updated. In [40], the bit energy metric is generated for SDM-NoC and therefore the overall energy could be evaluated with the updated value of E_{Rbit} and E_{Lbit}. For other NoC structures, the energy consumption can also be updated accordingly. The latency for the TDM-NoC is computed with three parts as in (15). For SDM-NoC, the first and second part remain the same which are independent of the NoC structure. However, the waiting time for congestion can be ignored in most cases because the physical links are shared and data is transported simultaneously for SDM-NoC which means that the required communication resources are likely to be reserved [39]. In this way, (15) can be simplified to (24) to calculate the latency for SDM-NoC.

$$L_{i,n} = (LC_{i,n} + \sum_{j=1}^{d_i^{SD}+1} LF_{L(j)})F_i^{SD}, \qquad (24)$$

As demonstrated above, the cost model CoREP not only can be utilized for SDM-NoC with minor modification, but also has the same scale of flexibility as for TDM-NoC which means that it is also applicable to most commonly used NoC topologies and routing algorithms for SDM-NoC because the evaluation of reliability cost, energy and latency are all topology and routing algorithm independent. To confirm these two points, a set of new experiments are designed and performed with four combinations of NoC topology and routing algorithm as shown in Table III. The benchmark HEVC is mapped onto these four SDM-NoCs. As shown in Fig. 16, optimal mapping patterns are found by the proposed approach for different SDM-NoC topologies and routing algorithms.

There are also emerging researches on the TDM-SDM combined NoC [41][42]. For this type of NoC structure, the reliability and performance can also be estimated straightforwardly similar to SDM-NoC. For energy consumption, the values for E_{Rbit} and E_{Lbit} need to be updated accordingly which is a potential topic for the studies for TDM-SDM combined NoC. Although no experiments have been done for this part, the flexibility of the proposed work is further enhanced.

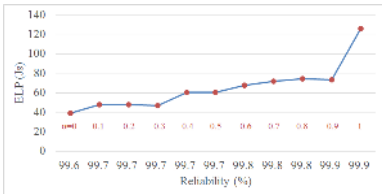


Fig. 14. Energy latency product (ELP) and reliability change with alpha changes for benchmark freqmine when $p_l = 0.01, p_h = 0.1$.

V. CONCLUSION AND FUTURE WORK

When dealing with the problem of searching for the best mapping pattern, it is of great importance to use appropriate

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2015.2422843, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems

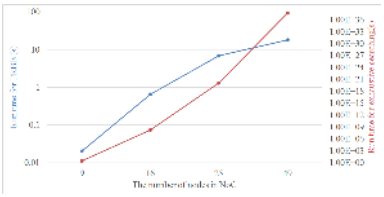


Fig. 15. Computation time of PRBB and exhaustive searching with respect to the number of nodes in NoC.

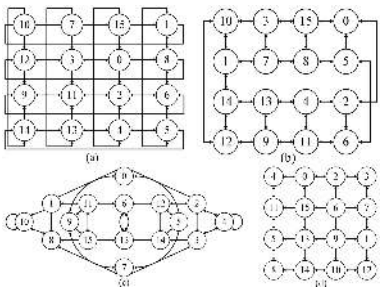


Fig. 16. Best mapping patterns of HEVC mapped onto four different combinations of topology and routing algorithm. The numbers means the index of IPs in the benchmark.

metrics for evaluation. In this paper, a highly efficient approach is proposed to co-optimize the reliability, energy and performance of a mapping pattern on NoC. Reliability, communication energy and performance are combined together by a general and highly flexible reliability-energy-performance formulation. A mapping approach is further proposed for finding the optimum mapping solution efficiently. Both theoretical analysis and experimental results show that this model outperforms classical and state-of-the-art approaches in terms of reliability, energy, latency, throughput and efficiency.

For further studies, the co-optimization during application mapping could be further improved by differentiating the importance of various measures. For the current model, it could be improved if separate weight parameters for energy and latency are defined respectively. In addition to reliability, communication energy and latency, other measures such as temperature could also be included for the future research which is a crucial factor for static energy consumption throughout the NoC. When the approach is extended to 3D NoC, thermal dissipation becomes an inevitable issue. The temperature model could also be used to further improve the model of link failure probability. Other influential factors such as impact from faulty component in the neighborhood could also be taken into account as future work.

REFERENCES

[1] J. B. Neil W. Bergmann, Sunil K. Shukla, "Quku: A dual-layer reconfigurable architecture," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 12, no. 1s:63, 2013.

[2] R. Marculescu, U. Ogras, L.-S. Peh, N. Jerger, and Y. Hoskote, "Outstanding research problems in NoC design: System, microarchitecture, and circuit perspectives," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 28, no. 1, pp. 3–21, 2009.

[3] B. Shim and N. Shanbhag, "Energy-efficient soft error-tolerant digital signal processing," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 14, no. 4, pp. 336–348, 2006.

[4] A. Kohler, G. Schley, and M. Radetzki, "Fault tolerant network on chip switching with graceful performance degradation," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 29, no. 6, pp. 883–896, 2010.

[5] J. Hu and R. Marculescu, "Energy- and performance-aware mapping for regular NoC architectures," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, no. 4, pp. 551–562, 2005.

[6] C.-L. Chou and R. Marculescu, "Farm: Fault-aware resource management in NoC-based multiprocessor platforms," in *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2011, pp. 1–6.

[7] A. Kahng, B. Li, L.-S. Peh, and K. Samadi, "ORion 2.0: A fast and accurate NoC power and area model for early-stage design space exploration," in *Design, Automation & Test in Europe Conference & Exhibition.*, 2009, pp. 423–428.

[8] J. Kim, M. Taylor, J. Miller, and D. Wentzlaff, "Energy characterization of a tiled architecture processor with on-chip networks," in *Low Power Electronics and Design, Proceedings of the International Symposium on*, 2003, pp. 424–427.

[9] J. Kim, D. Park, C. Nicopoulos, N. Vijaykrishnan, and C. Das, "Design and analysis of an NoC architecture from performance, reliability and energy perspective," in *Architecture for networking and communications systems, Symposium on*, 2005, pp. 173–182.

[10] M. Suess, R. Trautner, R. Vitulli, J. Ilstad, and D. Thurnes, "Technical dossier on on-board payload data processing," ESA, TECED-P/2011.110/Ms, 2011.

[11] T. R., "Esas roadmap for next generation payload data processors," in *Data Systems in Aerospace Conference (DASIA), Proceedings*, 2011, pp. 159–161.

[12] M. Shafto, M. Conroy, R. Doyle, E. Glaessgen, C. Kemp, J. LeMoigne, and L. Wang, "Modeling, simulation, information technology & processing roadmap," NASA, Technology Area 11, 2012.

[13] L. Ost and et al., "Power-aware dynamic mapping heuristics for noc-based mpsocs using a unified model-based approach," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 12, no. 3s:75, 2013.

[14] M. Farias, E. Barros, and A. Araujo, "An approach for multi-task and multi-application mapping onto NoC-based MPSoC," in *Circuits and Systems (MWSCAS), IEEE International Midwest Symposium on*, 2014, pp. 205–208.

[15] A. Banerjee, P. Wolkotte, R. Mullins, S. Moore, and G. Smit, "An energy and performance exploration of network-on-chip architectures," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 17, no. 3, pp. 319–329, 2009.

[16] D. Zhu, L. Chen, S. Yue, and M. Pedram, "Application mapping for express channel-based networks-on-chip," in *Design, Automation and Test in Europe Conference and Exhibition (DATE)*, 2014, pp. 1–6.

[17] S. Jafri, L. Guang, A. Hemani, K. Paul, J. Plosila, and H. Tenhunen, "Energy-aware fault-tolerant network-on-chips for addressing multiple traffic classes," in *Digital System Design (DSD), Euromicro Conference on*, 2012, pp. 242–249.

[18] C. Ababei, H. Kia, O. Yadav, and J. Hu, "Energy and reliability oriented mapping for regular networks-on-chip," in *Networks on Chip (NoCs), IEEE/ACM International Symposium on*, 2011, pp. 121–128.

[19] F. Khalili and H. Zarandi, "A reliability-aware multi-application mapping technique in networks-on-chip," in *Parallel, Distributed and Network-Based Processing (PDP), Euromicro International Conference on*, 2013, pp. 478–485.

[20] A. Das, A. Kumar, and B. Veeravalli, "Energy-aware communication and remapping of tasks for reliable multimedia multiprocessor systems," in *Parallel and Distributed Systems (ICPADS), IEEE International Conference on*, 2012, pp. 564–571.

[21] Z. Li, S. Li, X. Hua, H. Wu, and S. Ren, "Run-time reconfiguration to tolerate core failures for real-time embedded applications on NoC many-core platforms," in *High Performance Computing and Communications*, 2013, pp. 1990–1997.

[22] Y. Ren, L. Liu, S. Yin, J. Han, and S. Wei, "Efficient fault-tolerant topology reconfiguration using a maximum flow algorithm," *ACM Transactions on Reconfigurable TecACM Transactions on (TRETS)*, 2013, (In Press).

[23] M. Valinataj and S. Mohammadi, "A fault-aware, reconfigurable and adaptive routing algorithm for NoC applications," in *VLSI System on Chip Conference (VLSI-SoC), IEEE/IFIP*, 2010, pp. 13–18.

[24] B. T. William Dally, *Principles and practices of interconnection networks*. Morgan Kaufmann Publishers Inc., 2003.

[25] K. G. and P. D., "Dynamic power and thermal management of noc-based heterogeneous mpsocs," *ACM Transactions on Reconfigurable TecACM Transactions on (TRETS)*, vol. 7, no. 1s:1, 2014.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2015.2422843, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems

[26] M. Ni and S. Memik, "Thermal-induced leakage power optimization by redundant resource allocation," in *Computer-Aided Design, IEEE/ACM International Conference on*, 2006, pp. 297–302.

[27] T. Ye, L. Benini, and G. De Micheli, "Analysis of power consumption on switch fabrics in network routers," in *Design Automation Conference, Proceedings*, 2002, pp. 524–529.

[28] G. Bolch and et al., *Queueing networks and Markov chains: modeling and performance evaluation with computer science applications*. Wiley, 2006.

[29] S. H. Gerez, *Algorithms for VLSI design automation*. Wiley, 1998.

[30] S. Hollis, C. Jackson, P. Bogdan, and R. Marculescu, "Exploiting emergence in on-chip interconnects," *IEEE Transactions on Computer*, vol. 63, no. 3, pp. 570–582, 2014.

[31] [Online]. Available: http://www.carbondesignsystems.com/soc-designer-plus/

[32] [Online]. Available: http://venus.ece.ndsu.nodak.edu/ cris/software.html

[33] T. o. D. H. Erno Salminen, Ari Kulm ala, "Survey of network-on-chip proposals," *White Paper, OCP-IP*, pp. 1–13, 2008.

[34] P. Sahu and C. S., "A survey on application mapping strategies for network-on-chip design," *Journal of Systems Architecture (JSA)*, vol. 59, no. 1, pp. 60–76, 2013.

[35] T. Wiegand, G. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuit and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[36] G. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on Circuit and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[37] [Online]. Available: http://parsec.cs.princeton.edu/overview.htm

[38] J. T. Dertsimas, Dimitris, "Simulated annealing," *Statistical Science*, vol. 8, no. 1, pp. 10–15, 1993.

[39] A. Leroy, D. Milojevic, D. Verkest, F. Robert, and F. Catthoor, "Concepts and implementation of spatial division multiplexing for guaranteed throughput in networks-on-chip," *IEEE Transactions on Computer*, vol. 57, no. 9, pp. 1182–1195, 2008.

[40] S. H. Wang, A. Das, A. Kumar, and H. Corporaal, "Minimizing power consumption of spatial division based networks-on-chip using multi-path and frequency reduction," in *Digital System Design (DSD), 2012 15th Euromicro Conference on*, 2012, pp. 576–583.

[41] A. Lusala and J. Legat, "A hybrid NoC combining SDM-TDM based circuit-switching with packet-switching for real-time applications," in *New Circuits and Systems Conference (NEWCAS), 2012 IEEE 10th International*, 2012, pp. 17–20.

[42] A. Lusala and J. Legat, "A SDM-TDM based circuit-switched router for on-chip networks," in *Reconfigurable Communication-centric Systems-on-Chip (ReCoSoC), 2011 6th International Workshop on*, 2011, pp. 1–8.

Chen Wu received the B.S. degree from the School of Micro-Electronics and Solid-State Electronic from University of Electronic Science and Technology of China, Chengdu, China, in 2012. And he is currently a Master candidate in Institute of Microelectronics, Tsinghua University. His research interests include fault-tolerant system design, reliability modeling and reconfigurable network-on-chips.

Chenchen Deng received the B.S in Electronic Engineering from Beijing University of Posts and Telecommunications, China in 2007 and the DPhil in Engineering Science from University of Oxford, UK in 2012. She now works as a post-doctoral research fellow at Institute of Microelectronics, Tsinghua University, China. Her research interests include reconfigurable computing, 3-D IC and SoC design.

Leibo Liu received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 1999, and the Ph.D. degree with the Institute of Microelectronics, Tsinghua University, in 2004.

He is currently an Associate Professor with the Institute of Microelectronics, Tsinghua University. His current research interests include reconfigurable computing, mobile computing, and VLSI DSP.

Jie Han received the B.Sc. degree in electronic engineering from Tsinghua University, Beijing, China, in 1999 and the Ph.D. degree from Delft University of Technology, The Netherlands, in 2004. He is currently an assistant professor in the Department of Electrical and Computer Engineering at the University of Alberta, Edmonton, AB, Canada. His research interests include reliability, fault tolerance and energy efficiency, nanoelectronic circuits and systems, novel computational models for nanoscale and biological applications.

Jiqiang Chen received the B.S. degree from Department of Electronic Science and Technology from Xi'an Jiaotong University of China, Xi'an, China, 2012. He is currently a Master candidate in Institute of Microelectronics, Tsinghua University. His research interests include low-energy design, thermal management and reconfigurable network-on-chips.

Shouyi Yin received the B.S., M.S., and Ph.D degrees in electronic engineering from Tsinghua University, Beijing, China, in 2000, 2002, and 2005, respectively.

He was with Imperial College London, London, U.K., as a Research Associate. Currently, he is with the Institute of Microelectronics, Tsinghua University as an Associate Professor. He has published more than 20 referred papers, and served as a TPC member or reviewer for international key conferences and leading journals. His current research interests include mobile computing, wireless communications, and SoC design.

Shaojun Wei was born in Beijing, China, in 1958. He received the Ph.D. degree from Faculte Polytechnique de Mons, Mons, Belgium, in 1991.

He became a Professor with the Institute of Microelectronics, Tsinghua University, Beijing, China, in 1995. His current research interests include VLSI SoC design, EDA methodology, and communication ASIC design.

Prof. Wei is a Senior Member of the Chinese Institute of Electronics.