Making a Case for Partially Connected 3D NoC: NFIC versus TSV

AQEEB IQBAL ARKA, Washington State University, Pullman, WA

SRINIVASAN GOPAL, Intel Corporation, Hillsboro, OR

JANARDHAN RAO DOPPA, DEUKHYOUN HEO, and PARTHA PRATIM PANDE,
Washington State University, Pullman, WA

3D Network-on-Chip (3D NoC) enables design of high-performance and energy-efficient manycore computing platforms. Two of the commonly used vertical interconnection technologies are: through silicon via (TSV) and near-field inductive coupling (NFIC). Both TSV- and NFIC-based links introduce additional area overhead. One of the possible ways to reduce the area overhead is to design partially connected 3D NoC with minimal effect on overall performance. The achievable performance of the partially connected 3D NoCs depends on the area, energy, and bandwidth of the vertical links. Moreover, the electromigration-induced failure of TSV is more severe than the misalignment-induced error in NFIC. By considering all these pertinent factors, we demonstrate that it is indeed possible to design a partially connected NFIC-based 3D NoC that performs as good as a fully connected TSV-based counterpart when the data rate is below a certain limit. For higher data rates, the partially connected TSV-based 3D NoC is a viable solution. However, NFIC-based 3D NoC is always more robust than the TSV-based counterpart.

CCS Concepts: • **Networks** → **Network on chip**; • **Hardware** → *Radio frequency and wireless interconnect*;

Additional Key Words and Phrases: 3D Network-on-Chip (NoC), Near-Field Inductive Coupling (NFIC), Through Silicon Via (TSV), load balancing, multi objective optimization, MTTF, energy delay product (EDP)

ACM Reference format:

Aqeeb Iqbal Arka, Srinivasan Gopal, Janardhan Rao Doppa, Deukhyoun Heo, and Partha Pratim Pande. 2020. Making a Case for Partially Connected 3D NoC: NFIC versus TSV. *ACM J. Emerg. Technol. Comput. Syst.* 16, 4, Article 41 (August 2020), 17 pages.

<https://doi.org/10.1145/3394919>

1 INTRODUCTION

Near Field Inductive Coupling (NFIC) is a CMOS-compatible technology to design the vertical links in a 3D IC. Hence, NFIC can enable design of high-performance and energy-efficient 3D NoC architectures without any disruptive technology. Robustness is one of the most notable benefits of

This work was supported in part by the US National Science Foundation (NSF) grants CNS-1564014, CCF 1514269 and USA Army Research Office grant W911NF-17-1-0485.

Authors' addresses: A. I. Arka, J. R. Doppa, D. Heo, and P. P. Pande, Washington State University, 355 NE Spokane St, Pullman, WA 99163; emails: {aqeebiqbal.arka, jana.doppa, dheo, pande}@wsu.edu; S. Gopal, Intel Corporation, 2501 NE Century Blvd, Hillsboro, OR 97124; email: srinivasan.gopal@intel.com.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2020 Association for Computing Machinery.

1550-4832/2020/08-ART41 \$15.00

<https://doi.org/10.1145/3394919>

NFIC-based vertical links. The performance of an NFIC link is affected marginally by misalignment of inductor coils. We can maintain the target bit error rate (BER) by dissipating a modest amount of additional energy when two inductor coils are misaligned up to a certain limit. Current state-of-the-art research shows that inductively coupled vertical links can be expected to scale far more competitively with more advanced technology nodes than the through silicon via (TSV)-based counterparts [1]. These advantages can be achieved without incurring any additional fabrication cost that is inherent in TSV-based 3D integration.

There has been significant research on 3D NoC architectures. However, majority of 3D NoC architectures depend on TSV-based vertical links. TSVs suffer from bandwidth limitations due to their inherent low pass frequency response. They also experience large bandwidth reduction with just 10% misalignment errors [2]. TSVs are susceptible to electromigration (EM) and capacitive crosstalk-induced problems [3]. TSVs in a 3D NoC are placed in bundles to create the vertical links (VLs). In each VL, the crosstalk coupling increases the worst-case delay [4]. The combined effects of workload-induced stress and crosstalk lead to significant performance and reliability concerns for TSV-enabled 3D NoC architectures. Hence, the anticipated performance gain of 3D NoCs will be compromised due to TSV reliability issues. Inductor coils in NFIC-based VLs are also affected by EM-related issues as discussed in Reference [5]. However, by adjusting the inductor geometry by introducing reservoirs near vias and using thicker metal wires as proposed in Reference [5], it is possible to alleviate the effects of EM on NFIC-based links. Hence, the effect of workload-induced stress (EM) does not become a concern in NFIC-based VLs once the inductors are designed suitably. So, NFIC is an attractive alternative for designing 3D NoC architectures. However, the principal challenge with NFIC-based VLs is achieving suitable trade-off between data rate, energy, and area overheads. Therefore, it is essential to investigate performance-area trade-offs to make NFIC-based designs viable. To achieve high performance and energy-efficiency without introducing large area overhead, one possible solution is to restrict the number of allowable VLs. This in turn will make the 3D NoC partially connected. However, partially connected 3D NoCs introduce additional reliability challenges. Hence, it is extremely important to adopt a suitable design optimization strategy specifically for partially connected 3D NoCs. The design and associated optimization strategies should depend on the characteristics of the physical layer (data rate, area, and energy). In this article, our aim is to undertake a comprehensive exploration of design and optimization methodologies for partially connected 3D NoCs. We demonstrate that, depending on the characteristics of the physical layer, partially connected 3D NoC is a viable solution. By adopting a suitable design optimization strategy, a partially connected 3D NoC achieves an appropriate balance between overall performance and area overhead.

The rest of the article is organized as follows: The related work is discussed in Section 2. Then, in Section 3, we discuss the challenges of NFIC- and TSV-based NoC design. In Section 4, we present our problem setup and the proposed design optimization methodology. In Section 5, we present the experimental results; Section 6 concludes the article.

2 RELATED WORK

TSV is the most prevalent technology for designing VLs in 3D NoCs. Generally, multiple TSVs are combined to design a VL with higher aggregate data rate [6]. Several studies have been conducted on 3D NoC architectures based on TSV links. Most of the existing works on TSV-based architectures have focused on the performance benefits of 3D integration using TSVs [7–9]. EM-induced TSV failures and modeling of TSVs under EM [10, 11] have been studied, too. The effects of workload-induced stress on the TSV-based VLs have also been addressed [12]. The combined effects of EM and crosstalk on Mean Time to Failure (MTTF) and reliability of TSV-based 3D NoC have been studied [13]. It has been shown that a small-world network-enabled 3D NoC

(3D SWNoC) outperforms mesh-based regular architectures and also other application-specific irregular architectures [7, 14].

Wireless 3D technology using proximity communication has been an alternative vertical integration technique capable of demonstrating low bit error rate (BER) at high speed and low power using inductive and capacitive coupling [15,16]. Chip-to-chip data transmission in non-return-to-zero (NRZ) format can be directly accomplished using wireless channels formed by capacitive coupling (CC) or NFIC without any carrier modulation for vertical communication. CC is a face-to-face wireless 3D integration technique, applicable for ultra-short-reach communications using the metal plates of the top layer. The capacitance between the metal plates is employed for channel formation. The channel capacitance directly dictates the gain. In CC, the receiver directly detects the transmitting signal's voltage transitions to recover NRZ data. The CC channel is more often restricted to face-to-face 3D integration [16].

NFIC is capable of achieving low bit error rate (BER) with high bandwidth and low power consumption [16]. NFIC-based VLs can be designed using the top metal layer of each tier [17]. Hence, they do not incur any additional fabrication cost [18]. This makes NFIC-based link an attractive alternative to TSV for 3D NoC design. Furthermore, the resiliency of NFIC-based links in presence of misalignment has also been studied, where the communication capability of the link can be preserved with the expense of modest amount of additional energy consumption [16, 19]. The design and implementation of high data rate NFIC link have been discussed in a previous work [20]. However, the advantages of NFIC-based links come at the cost of area and energy overhead [19]. Advantages and limitations of NFIC-enabled 3D NoC architectures have been explored in several previous works. These include the use of NFIC bypass links to connect non-adjacent tiers of a 3D NoC [19, 21, 22] and use of NFIC-based links as randomized shortcuts to enhance NoC performance [23]. Further work also investigated the application of NFIC-based links in irregular NoC design including NoCs with a "staircase" or a "checkerboard" layout [24]. There is also a study on 3D NoC containing both TSVs and NFICs as VLs [19]. The reliability of the resultant hybrid NoC was shown to increase significantly even with a limited number of NFIC-based links compared to an exclusive TSV-based 3D NoC [19].

For partially connected 3D NoCs, "Elevator-first" algorithm has been proposed [25] as a deadlock-free distributed routing method. Further studies have shown that Adaptive Layered Shortest Path (ALASH) routing can outperform elevator-first for irregular topologies [26]. Hence, for irregular topologies like small-world network-based NoC architecture ALASH is the preferred routing method.

There is an important body of work in NoC literature that deals with NoC traffic characteristics and unevenness of the workload [27]. Prior works also explored application of information theoretic approach like network coding to enhance performance of NoC architectures [28, 29]. Our work complements these efforts by studying effects of traffic distribution on vertical link reliability in 3D NoC architectures.

It is important to note that none of the previous works include the comprehensive design and optimization of a partially connected NoC based on physical layer characteristics of TSV- and NFIC-based links. Hence, our work complements the existing body of literature by presenting a thorough analysis of partially connected 3D NoC architectures with TSV- and NFIC-based links.

3 CHALLENGES OF NFIC- AND TSV-BASED NOC DESIGN

In this section, we briefly summarize the challenges associated with designing 3D NoC architectures using both TSV- and NFIC-based VLs. The principal issues are the performance-area trade-offs and reliability (hence MTTF). We also discuss the need for load balancing for optimizing NoC design and present our rationale behind designing a partially connected 3D NoC.

Among the state-of-the-art 3D integration technologies, TSV is the most mature and popular technology. NFIC-based VL is an attractive alternative to TSV-based counterpart. NFIC-based links are contactless where the communication happens through the mutual inductance of two inductor coils placed in different layers of the 3D IC. However, the design of an energy-efficient NFIC link with small area overhead is challenging. The effectiveness of NFIC links depends on the electromagnetic coupling between the transmitter and receiver coils. Higher data rate would cause increased current swing in the NFIC links, which results in higher energy consumption [19]. So, designing an optimum NFIC-link depends on the suitable trade-off between area overhead and energy consumption.

There is another issue that may arise when designing NoC primarily based on NFIC-based VLs—that is the problem of neighboring chip interference. The case of neighboring chip interference along the same pillar (same X-Y coordinate) is a valid concern for a fully connected 3D NoC using only NFIC-based VL. In a neighboring vertical chip-to-chip communication, interference occurs if there exists a vertical link between 1st and 2nd chips and another vertical link simultaneously exists between the 3rd and the 4th chips. In this specific case, an unintentional link is formed between the 2nd and 3rd chip due to the vertical broadcast nature of electromagnetic flux in NFIC links resulting in interference. To prevent interference, metal shields aligned between inductors in links, which receive unintentional electromagnetic flux, are used [30]. Electromagnetic flux is reduced by eddy-currents created in shields. As a result, the coupling coefficient between the inductors for the unintentional link is reduced and thus it reduces the crosstalk voltage. However, the shield introduction comes with an additional power overhead in the intentional link [30]. Staggered stacking is another approach to prevent unintentional crosstalk [31]. All in all, the issue with interference can be taken care of with appropriate physical design of the NFIC-based VLs. In this work, we have assumed that the physical design has been done accordingly to alleviate all such issues of the VLs before implementing an NFIC-based NoC.

The complexity of designing a low energy NFIC link results in a link with a larger area overhead compared to its TSV-based counterpart. It is possible to decrease area and increase the data rate of the NFIC-based links at the expense of energy consumption. However, NFIC-based links can tolerate significant amount of misalignment between the central axes of the transmitter and the receiver inductor coils [19]. Higher communication range for NFIC links requires increased inductor diameter, which leads to larger link area [19]. All these factors make the performance-area trade-off very important for NFIC-based NoC design.

There are significant challenges associated with NoC design using TSV-based links as well, because TSVs suffer from some serious reliability issues [32]. The precise mechanical alignment requirement during the bonding and stacking of planar dies results in reduction of available TSV bandwidth. It also introduces bonding faults, and in the worst case, causes failure of the links resulting in low yield and high manufacturing cost [33]. EM due to workload-induced stress is another factor behind the failure of TSV-based links. In a TSV-based 3D NoC, the VLs undergo a large variation of workload distribution depending on the architecture and application [12]. This workload-induced stress causes the TSV-based VLs to experience EM on the TSV junctions and landing pads. Higher workload intensity leads to more severe EM effects. EM results in the creation of a void at the TSV landing pad, which increases the resistance. The effect of EM-induced resistance increase is quantified by the following equation:

$$R(t) - R_o = A \ln \frac{t}{t_o}, \text{ for } t > t_o. \quad (1)$$

Here, R_o is the initial TSV resistance, which mainly depends on the fabrication process and component materials. The parameter t_o is the time when the size of void becomes equal to or greater than the cross-sectional area of the TSV. The aging coefficient A depends on the amount

of current that passes through TSVs in unit time. In this work the value of A was taken to be $0.05 \text{ k}\Omega/\log(s)$ based on the analysis reported in prior work [3].

As a result of the increase in resistance, the delay of the TSV increases causing the TSV to fail. In general, 10% increase in delay is considered as failure of a TSV [3, 11]. We follow the same assumption in this article. Higher utilization would lead to lower MTTF for each TSV. Eventually, failure of TSV links will result in performance degradation and lifetime reduction of NoC. Another cause for delay increment in TSVs is the crosstalk capacitance. Crosstalk noise in a TSV bundle depends on the exact TSV location (which determines the proximity to other TSVs in the bundle) and the bit pattern.

For TSV-based designs, multiple TSVs are placed in a bundle and combined together to achieve high data rate [6]. Multiple TSVs in a bundle introduce area overhead due to pitch and keep-out-zone (KOZ) requirements for obtaining high yield [34]. Hence, the large area overhead associated with the pitch and keep-out-zone of a TSV-bundle is a concern [10]. This large area overhead should be considered while designing an NoC with TSV-based VLs.

The area overheads associated with both TSV- and NFIC-based links motivate us to explore design and associated optimization methodologies for partially connected 3D NoCs. Area savings in partially connected 3D NoCs should not come at the cost of significant performance degradation. The VLs enable performance gain in a 3D NoC as long-range shortcuts [35]. Hence, the placement of the VLs in a partially connected 3D NoC is an important design optimization challenge that needs to be thoroughly investigated.

4 PROBLEM SETUP AND SOLUTION APPROACH

In this section, we present an optimization framework necessary to design a partially connected 3D NoC architecture. Following the proposed optimization methodology, we establish the performance-area trade-offs for both TSV- and NFIC-enabled partially connected 3D NoCs.

4.1 Problem Description

To carry out the design and optimization for partially connected 3D NoC, first, we need to consider a suitable interconnection architecture. We focus our analysis on the small-world network-enabled 3D NoC (SWNoC), as it has been shown to outperform other regular or irregular 3D NoC architectures [7, 14]. In a 3D SWNoC, the VLs establish the long-range shortcuts necessary in a small-world network and they are responsible for the performance gain. However, prior work on 3D SWNoC design [12] has shown that there is uneven load distribution (NoC traffic) among the VLs, which causes workload-induced stress and gives rise to permanent failures. This leads to the reduction in overall lifetime of the NoC. This is specifically applicable for TSV-based designs. By attempting to balance the load across the VLs, it is possible to improve the overall throughput of the NoC resulting in improved performance [36]. Load-balancing the NoC by adjusting link and tile placement tries to bring highly communicating tiles closer together and attempts to improve throughput by utilizing the given resources more efficiently. This reduces uneven loads across multiple VLs that in turn increases the lifetime of VLs to enhance the reliability of the complete NoC. Even after balancing the load across VLs, utilization of some VLs will still be significantly lower than the others. Hence, removing these lowly utilized VLs is a possible strategy to save substantial area overhead without introducing significant impact on performance. Reducing the number of VLs will affect the performance of the SWNoC. To keep the total number of links fixed, we add additional planar links while reducing the number of VLs. However, we need to consider an optimized placement of both planar links and VLs such that there is negligible impact on performance. We need to consider these constraints for both TSV- and NFIC-based partially connected SWNoC architectures.

Table 1. Physical Parameters of NFIC- and TSV- based Vertical Links [19, 45]

Data Rate (Gbps)	Area of NFIC	TSV Bundle structure(X*Y)	Area of TSV Bundles	Area ratio(A_{NFIC}/A_{TSV})	Link Energy NFIC (pJ)	Link Energy TSV (pJ)
18	1,575	1*3	1,368	1,151	7.36	7.26
21	1,332.7	2*2	1,968	0.677	15.84	6.72
24	1,267.5	2*2	1,968	0.644	33.504	7.04

Our aim is to determine the SWNoC architecture with the following constraints:

$$N_{NFIC} \cdot A_{NFIC} \approx N_{TSV} \cdot A_{TSV}. \quad (2)$$

Where, N_{NFIC} : Number of NFIC links; A_{NFIC} : Area of each NFIC link; N_{TSV} : Number of TSV Links; and A_{TSV} : Area of each TSV Link. This constraint allows us to design a partially connected NFIC-based SWNoC with equal area overhead as a fully connected TSV-based counterpart and vice versa.

We present different configurations of NFIC- and TSV-based links in Table 1. Table 1 shows that, for lower data rates (18 Gbps): $A_{NFIC} > A_{TSV}$, which leads to the motivation for a partially vertically connected NoC designed with NFIC links ($N_{NFIC} < N_{TSV}$). Conversely, for higher data rates (21 and 24 Gbps) TSV-based links have higher area overhead ($A_{NFIC} < A_{TSV}$), which necessitates the design of a partially vertically connected TSV-based NoC ($N_{NFIC} > N_{TSV}$). Hence, we set an upper bound on the number of VLs in a partially connected NoC using Equation (2). With this constraint, we need to evaluate the overall performance of the partially connected 3D NoCs considered in this work.

4.2 NoC Design Optimization

There are two ways to achieve performance and reliability trade-off in any NoC. We can either focus on load balancing among the NoC links or we can minimize the energy-delay-product (EDP). Both optimization objectives have associated advantages and limitations. If we balance the load across the NoC links, then the overall throughput will be maximized, too. This improvement in throughput happens, as load balancing makes some links more readily available for utilization, thus decreasing the congestion [36]. This makes the overall utilization of each VL in the NoC more evenly balanced. Eventually, this will help in improving the lifetime of NoC-enabled many-core system [12]. Overall, load balancing increases the NoC throughput while also reducing the excessive traffic through highly utilized links that helps in improving MTTF and hence, reliability. On the contrary, if we solely focus on maximizing NoC performance, we should aim to minimize the NoC latency and energy consumption. This minimizes NoC energy-delay-product (EDP), which is a combined metric considering both latency and energy. While minimizing the EDP, there will be VLs that are utilized more than the others (as mentioned earlier). This of course improves the performance but can potentially make very highly utilized VLs susceptible to early failure, hence compromising the lifetime of the manycore system [12, 14].

Load Balancing-based Optimization (LB): The load-balancing approach attempts to improve the overall throughput by utilizing the given resources more efficiently. To balance the expected VL utilization (which also load-balances the NoC), we consider minimizing both the mean ($\bar{U}$) and standard deviation (σ) of expected VL utilization as suitable objectives.

The expected utilization of k th link (u_k) can be obtained by the following equation:

$$u_k = \sum_{i=1}^N \sum_{j=1}^N (f_{ij}(t) \cdot p_{ijk}). \quad (3)$$

Here, N is the total number of cores and p_{ijk} indicates whether a link k is used to communicate between core i and core j , respectively, i.e.,

$$p_{ijk} = \begin{cases} 1, & \text{if cores } i, j \text{ communicate along link } k \\ 0, & \text{otherwise} \end{cases}.$$

p_{ijk} can be determined by using the NoC connectivity and the adopted routing protocol, $f_{ij}(t)$ is the communication pattern as a function of time and includes the temporal variations of the workload. We capture the $f_{ij}(t)$ parameter directly from detailed Gem5 simulations by creating checkpoints [37] after specific intervals of the application execution. It should be noted that creating checkpoints during a full-system simulation in Gem5 does not hamper the execution but allows us to study workload statistics over small periods of time. The overall application behavior is reconstructed by combining the $f_{ij}(t)$ observed over these smaller time periods.

As a consequence, the mean ($\bar{U}$) and standard deviation (σ) of link utilization can be determined from the following equations:

$$\bar{U} = \frac{1}{L} \sum_{k=1}^L u_k, \quad (4)$$

$$\sigma = \sqrt{\frac{1}{L} \sum_{k=1}^L (u_k - \bar{U})^2}. \quad (5)$$

Here, L is the total number of links in the NoC.

However, considering only the throughput of NoC design is not enough to encapsulate the overall performance of the NoC. Energy is another very important metric to quantify the performance of the NoC. The energy consumption of a candidate NoC design consists of energy dissipated by both links and routers, i.e.,

$$E = \sum_{i=1}^L E_{link,i} + \sum_{j=1}^N E_{router,j}. \quad (6)$$

Where, E_{link} and E_{router} refer to the energy consumed by the links and routers, respectively; and L is the total number of links and N is the total number of routers in the NoC. It should be noted that each core is connected to a router and hence the number of routers is equal to the number of cores.

To capture the effect of both throughput and energy on the NoC performance, we need to adopt a multi-objective optimization (MOO) methodology for the NoC design. We can write our combined objective function for the MOO-based optimization as follows:

$$D^* = MOO(D, OBJ = \{\bar{U}(d), \sigma(d), E(d)\}). \quad (7)$$

Where, D^* is the set of Pareto-optimal designs among all possible NoC configurations D , i.e., $D^* \subseteq D$; MOO is a multi-objective optimization solver; and OBJ is the set of all objectives to evaluate a candidate NoC design $d \in D$. For TSV-based NoC design, adopting load balancing gives rise to more even VL utilization compared to performance optimization exclusively [11]. This even utilization leads to improvement in lifetime of TSV-enabled VLs, thus enhancing reliability.

Performance-based Optimization (PO): For the EDP-based optimization, we have to consider the communication cost (C) and energy as the two objectives. The communication cost is defined in Equation (8) below. Here, C for a candidate NoC design is defined as follows:

$$C = \left(\sum_{i=1}^N \sum_{j=1, i \neq j}^N (R_{delay} * h_{ij} + d_{ij}) * f_{ij}(t) \right). \quad (8)$$

Where, $f_{ij}(t)$ is the same as Equation (3) and d_{ij} is defined as the total inter-router delay between the i th and the j th nodes, h_{ij} is the hop count between i th and j th nodes, and R_{delay} denotes the intra-router delay of a single router. The intra-router delay parameter (R_{delay}) includes the effect of router size and pipelining, where each pipeline stage is constrained to have a delay of 15 fan-out-of-four (FO4) [38]. The communication cost actually represents the overall NoC latency. Therefore, the multi-objective optimization formulation for this case becomes

$$D^* = MOO(D, OBJ = \{E(d), C(d)\}). \quad (9)$$

Here, we use the energy and latency in a multi-objective optimization setup to capture the joint effect of energy-delay product (EDP) to consider the effect of both NoC latency and energy together. This optimization formulation is specifically applicable for NoC design with NFIC-based VLs. NFIC-based VLs are contactless links. Hence, with properly designed robust inductor coils [5], they do not suffer from workload-induced EM. Highly utilized NFIC links do not fail due to workload-induced stress. Hence, instead of load balancing, we focus on minimizing EDP.

Optimization Methodology: To ensure that the proposed NoC design is plausible and energy-efficient, we set the following constraints over the design space:

1. The NoC must be a connected graph so that every source and destination pair can communicate with each other.
2. The maximum number of ports (k_{max}) should be less than an allowable upper limit ($k_{max, allowed}$) to avoid large routers that tend to be power-hungry. It has been shown experimentally that for a 3D SWNoC, $k_{max, allowed}$ is 8 [39]. In other words,

$$k_{max} < k_{max, allowed}.$$

3. The number of links in a 3D SWNoC ($L_{3D\ SWNoC}$) should be equal to that of a 3D MESH NoC ($L_{3D\ MESH}$) of the same system size, i.e.:

$$L_{3D\ SWNoC} = L_{3D\ MESH}.$$

This constraint ensures that the 3D SWNoC does not introduce any additional links. It should be noted that, when creating a partially connected 3D NoC, we remove VLs but the total number of links in the NoC remains the same as more planar links are added.

Finally, we formulate the NoC design as a MOO problem as follows:

$$\begin{aligned} D^* &= MOO(D, OBJ) \\ OBJ &= \begin{cases} \{\bar{U}(d), \sigma(d), E(d)\}, & \text{for LB optimization} \\ \{E(d), C(d)\}, & \text{for PO optimization} \end{cases} \end{aligned}$$

Here, D , D^* , and OBJ are same as in Equation (7).

Figure 1 shows the overall flow of the optimization methodology for the SWNoC design. As a first step, we design fully connected TSV- and NFIC-based SWNoC architectures following the MOO scheme described above. This part of the performance evaluation sets up the upper limits for both achievable performance, area overhead, and reliability.

For the partially connected 3D NoC designs, we reduce the number of VLs. Hence, we would include the constraint provided by Equation (2) to the above-mentioned optimization formulation. We take the optimized fully connected 3D NoC design created from the first step of the optimization methodology and then remove the relatively lowly utilized links. While reducing the number of VLs, we still maintain the total number of links (planar + vertical) to be unchanged. Hence, this gives us opportunity to add more planar links in the SWNoC architecture. In the 2nd step, we optimize the placement of the remaining VLs and the newly added planar links.

In a MOO problem, it is challenging to identify the best solution that performs equally well for all the specified objective functions. Exploring the NoC design space to look for the best possible

Fig. 1. Illustration of the overall flow of optimization methodology for designing fully and partially connected 3D SWNoC using different set of objectives via AMOSA algorithm. The 2nd step of the optimization process derives the partially connected architecture from the optimized fully connected 3D NoC (1st step).

solution is computationally expensive and sub-optimal in terms of the uncovered designs in a reasonable amount of time. To solve this Multi-objective optimization problem, we use Archived Multi Objective Simulated Annealing (AMOSA) [40]. AMOSA is based on simulated annealing and is guided by a set of archived solutions and is known to outperform many existing MOO solvers. During each optimization step there are two possible perturbations that we can perform to the existing NoC design configuration to move to the new state (successor function).

- *New Link Placement*: Select and remove a link between two randomly selected routers.
- *Core Swap*: Interchange the position of two random cores.

Depending on the Pareto dominance condition of the new configurations and the current solution, AMOSA updates its archives and continues till convergence. Thus, we obtain a set of archived candidates NoC configurations D^* . From those candidate NoC designs, we choose the configuration with the best performance (in terms of EDP). In summary, we employ AMOSA for both load balancing and EDP optimization. This gives rise to optimized NoC configurations commensurate with our design objectives.

The convergence of AMOSA is tested as follows: We employed a maximum of 1K local search iterations. After each iteration, we keep track of the updated Pareto set of designs. If this Pareto set does not change in 100 consecutive iterations, we declare convergence. Specifically, we can track the Pareto hyper-volume (PHV)—metric to measure the quality of Pareto set—over local search iterations and if this result does not change in 100 consecutive iterations, then we declare convergence. In practice, we observed that AMOSA converges in ~ 200 iterations. The overall runtime depends on the number of objectives in our multi-objective optimization formulation. For two objectives, AMOSA converged in 40 hours. For three objectives, AMOSA took 70 hours to converge. These wall-clock times are based on experiments run on a machine with the configuration: Intel® Xeon® CPU E5-2630 @ 2.2 GHz machine with 252 GB RAM running CentOS 6.

5 RESULTS AND ANALYSIS

In this section, our aim is to quantify the performance-area-reliability trade-off of both NFIC- and TSV-based NoC architectures. We consider both partially and fully connected configurations in the overall performance analysis. We present the performance-reliability trade-off for both LB and PO NoC optimization strategies.

5.1 Experimental Setup

To evaluate the performance of different NoCs, we use Gem5, a full system simulator, and obtain detailed network level information [37]. To implement the different NoCs considered in this work,

Fig. 2. MTTF distribution of TSV-based VLs for different benchmarks (a) VIPS (b) CAN, and (c) RAD.

we modified the Garnet network in Gem5 [41]. We use the MOESI_CMP_directory cache coherence protocol. Each core consists of private 64 KB L1 instruction and data caches and a shared 8 MB L2 cache. Our system consists of 64 cores partitioned into four layers. The routers are synthesized from an RTL-level design using TSMC 28-nm CMOS process in SynopsysTM Design Vision. Each router port has a buffer depth of two flits and each router port has four virtual channels. For an irregular architecture like the SWNoC, the topology-agnostic ALASH algorithm is adopted [42]. ALASH algorithm is also used for partially connected NoC configurations, as it has been shown to outperform elevator-first routing for irregular topologies [26]. The energy consumption of the NoC routers was obtained from the synthesized netlist by running SynopsysTM Prime Power, while the energy dissipated by wireline links was obtained through HSPICE simulations considering the length of the wireline links. We consider four SPLASH-2 benchmarks, namely, FFT, RADIX(RAD), LU, and WATER(WAT) [43], and four PARSEC benchmarks, DEDUP(DED), VIPS, FLUIDANIMATE (FLU), and CANNEAL(CAN) [44] in this performance evaluation. For each type of vertical link, the energy and delay profiles of the link were incorporated in the NoC simulator to evaluate the overall performance.

5.2 Reliability-performance Trade-offs due to Optimization Objectives

In this section, we discuss the effects of workload-induced failure in TSV-based 3D NoC. We also discuss the effect of considering different optimization objectives on the reliability and performance of the NoC.

In Section 3, we discussed the effects of workload-induced stress on TSV reliability. Failure of TSV-based VLs compromises the overall performance of the NoC [12]. Failure of VLs causes the utilization of the remaining links to increase even more, decreasing the Mean Time to Failure (MTTF). We can address this situation by targeting load balancing as the optimization goal instead of sole performance (EDP). It makes the VLs more evenly utilized. Hence, utilization of highly utilized links reduces and thus the workload-induced stress reduces. As an example, and for brevity, we chose 21 Gbps TSV-based VLs for the following analysis. The effect of workload-induced stress on TSV MTTF for different NoCs is illustrated in Figure 2. Here, PO NoC denotes Performance optimized NoC, and LB NoC denotes Load Balanced NoC. Figure 2 shows the frequency distribution of VLs over the normalized MTTF. Figures 2(a), (b), and (c) show the MTTF distributions for different NoCs for VIPS, CAN, and RAD benchmarks, respectively. When comparing the PO and LB NoCs, the MTTF distribution moves to the right for Load Balanced (LB) architecture. This shows that load balancing improves the MTTF of VLs for TSV-based NoCs compared to the MTTF distribution of PO TSV-based NoCs. It can be seen from Figure 2(a) that in VIPS benchmark a total of 43 VLs fail before 30K hours (approximately 3.5 years) of operation in

Fig. 3. Performance-reliability trade-off between Performance Optimized (PO) and Load Balanced (LB) TSV-based 3D NoC.

Fig. 4. Performance comparison between Load Balanced (LB) vs. PO NFIC-based 3D NoC.

PO NoC. However, load balancing vastly improves the scenario by providing MTTF improvement of over 25%. A similar trend can be seen in Figure 2(c) for the RAD benchmark, where 37 VLs fail cumulatively after 30K hours. Noticeably, in the CAN benchmark, the links are more evenly utilized in all the cases, which leads to much better MTTF compared to the other benchmarks (approximately 30K hours of operation before 11 VLs fail in total). However, load balancing still improves MTTF over the PO NoC (only 5 VLs fail after 30K hours of operation).

However, the lifetime improvement from load balancing comes at the cost of performance degradation. Load balancing attempts to redistribute the traffic over the NoC to even out utilization. This causes hop count to increase for some messages, which degrades the performance. Despite the performance drop, there is a clear reliability improvement, which is evident in Figure 2. Figure 3 illustrates the performance and reliability trade-off between PO and LB TSV-based NoC. We can see that there is an average performance degradation of 9% with MTTF improvement of over 14%. Here, it is clear that the improvement in MTTF is more than the performance degradation. This validates the need of LB optimization in case of TSV-based NoC (where EM is an important bottleneck).

On contrary, workload-induced EM is not an issue for NFIC-based VLs when the inductor coils have been designed with reservoirs near vias as discussed in Reference [5]. So, the MTTF enhancement offered by LB optimization is not necessary for 3D NoC based on NFIC-based VLs. The performance comparison between PO and LB NFIC-based 3D NoC should be the motivating factor of choosing the optimization metric that should be considered for NFIC-based NoC. Similar for TSV-based VLs, we chose 21 Gbps NFIC-based VLs as an example physical layer. Figure 4 demonstrates the EDP comparison between LB and PO NFIC-based NoC architectures, where the EDP is normalized with respect to the PO NFIC-based NoC. There is a performance degradation between 7%–12%, depending on the benchmark. However, there is no apparent improvement in reliability, as workload-induced stress is not a concern in this case. This result attests to the fact that sole performance optimization should be the target for NFIC-based NoCs.

5.3 Performance-area Trade-off of TSV- and NFIC-based 3D NoC

In this section, we analyze the performance area trade-off for different 3D NoC architectures based on the physical layer characteristics of the VLs. As discussed in the previous sub-section, the performance metric to be optimized also depends on the physical layer. For TSV-based NoCs, we should use LB NoC design. However, we would choose PO NoC design for NFIC-based architectures. However, as mentioned above in Equation (2), our target is to make the area overheads of both TSV- and NFIC-enabled NoCs equal. As the area overhead varies based on the characteristics of the physical layer, we investigate the possibilities of designing both fully and partially connected architectures.

Fig. 5. Performance for CAN benchmark: (a) For NoCs based on VLs with data rate less than or equal to 18 Gbps; (b) For NoCs based on VLs with data rate greater than 18 Gbps.

As we can see from Table 1, NFIC link with 18 Gbps data rate has a significantly larger area than their TSV-based counterpart, whereas for 21 and 24 Gbps, NFIC-based VLs are smaller than TSV-based VLs. This leads us to explore performance-area trade-offs of partially connected NFIC-based 3D NoC for VLs with data rate of 18 Gbps. By the same token, we should consider partially connected TSV-based 3D NoC for VLs with 21 and 24 Gbps data rates. These configurations help us to meet the constraint of area overhead set up by Equation (2). However, we need to analyze the impacts of partial connection on the overall performance of the NoC. In this regard, the energy consumption of NFIC- and TSV-based VLs is an important parameter. For 18 Gbps links, the energy consumptions of both the links are very close. However, for 21 and 24 Gbps data rates, NFIC consumes more energy than the TSV-based counterpart. NFIC-based VLs have to drive higher current through the coils to achieve enhanced data rates resulting in higher power consumption.

In Figure 5, we compare the performance of NoCs based on different VLs for the CAN benchmark as an example. The comparison of TSV- and NFIC-based architectures is divided in two groups: VLs with data rate of 18 Gbps where area overheads make partially connected NFIC-based 3D NoC necessary; the second group includes VLs with data rates of 21 and 24 Gbps where partially connected TSV-based NoC should be considered. It should be noted that we are undertaking this analysis based on the physical layer characteristics that we obtained through our design and implementation. For a different implementation of the physical layer, we may have different area constraints. However, the methodology adopted here is valid irrespective of the physical layer characteristics. In Figure 5, we illustrate the latency, energy, EDP, and area of different NoC architectures of the two groups. All the performance metrics such as latency, energy, and EDP in Figure 5 are normalized with respect to the fully connected TSV-based NoC of the same VL data rate. From Figure 5, we can infer that the latency of NFIC-based NoCs is consistently better for all data rates. This is due to the PO-based approach of NoC design, which helps to reduce hop count and improve latency. However, the energy consumption of NFIC-based NoC is higher in all the cases except for the 18 Gbps data rate. This trend is especially apparent for 24 Gbps where the energy consumption for NFIC VLs is five times that of TSV VLs.

In case of 18 Gbps links, fully connected NFIC-based NoCs have better EDP than the corresponding TSV-based counterparts. However, as discussed above, partially connected NFIC-based NoCs have to be considered here due to the higher area overhead. To match the area overhead constraint shown in Equation (2), the number of allowable VLs in the NFIC-based NoC has to be reduced to 41 VLs instead of the regular 48 VLs for a fully connected system with 64 cores. These partially connected NFIC-based NoCs can match the EDP of fully connected TSV-based NoC without the area overhead, as seen in Figure 5(a).

Fig. 6. Performance variation with NoC with Same Vertical Link Area, all numbers normalized w.r.t. fully connected TSV-based NoC.

In the case of VLs with data rates of 21 and 24 Gbps, we consider partially connected TSV-based NoCs. To meet the area constraint, the number of allowable VLs for 21 and 24 Gbps TSV-based NoCs are 32 and 30, respectively. When we consider 21 Gbps links, fully connected NFIC-based NoCs are only slightly worse than fully connected TSV-based architecture with lower area overhead. This is due to the combined effect of higher energy consumption of VLs and lower latency as a result of the PO-based approach. The partially connected TSV-based NoC has the same area as the fully connected NFIC-based NoC in this case. This happens at the cost of reliability. Only for high data rate NFIC link's (24 Gbps) high energy overhead causes EDP to be much worse compared to both fully and partially connected TSV-based counterparts.

In Figure 6, we compare the performance of NoCs with the same area overhead but based on different type of VLs. For data rate of 18 Gbps, we have fully connected TSV-based NoC and partially connected NFIC-based one. For 21 and 24 Gbps, we have the opposite configuration. Still, in all configurations the area overheads of NFIC- and TSV-based NoCs are equal. For 18 Gbps VLs, partially connected NFIC-based NoCs match the EDP of fully connected TSV-based counterparts. Here, there is an average EDP improvement of 4% for the partially connected NFIC-based NoC. In case of 21 Gbps VLs, a fully connected NFIC-based architecture has similar area of a partially connected TSV-based NoC while having equivalent or better EDP. In this case the fully connected NFIC-based NoC has 2% better performance on an average when compared to a partially connected TSV-based NoC. However, for 24 Gbps VLs, NFIC-based NoCs always have worse EDP than their TSV-based counterparts in all the cases. On an average, for 24 Gbps links, the fully connected NFIC-based NoC has 20% more EDP than its partially connected TSV-based counterpart.

In Figure 7, we compare the performance and associated area overheads of fully and partially connected NoCs based on both TSV- and NFIC-based VLs. Here, in all the cases, the performance is slightly degraded when going from a fully connected to a partially connected NoC, but the area gains are significant. This degradation is due to the fact that removing VLs increases hop counts and latency. Utilizing the remaining VLs more causes an increase in energy consumption. However, in case of TSV-based NoCs, changing the architecture from fully connected to a partially connected also results in a significant decrease in MTTF due to the higher traffic through the remaining VLs. This effect is shown in Figure 8 for a high injection benchmark (RAD) as an example. As seen in Figure 8, compared to a fully connected TSV-based NoC, the MTTF distribution of a partially connected counterpart shifts to the left as expected. This indicates lower reliability for the partially connected NoC, as 16 VLs out of 29 fail before 18K hours of operation (approximately two years). However, the benefits of LB optimization become even more apparent in this case, as only 5 VLs

Fig. 7. Performance & area variation with NoC with same vertical link type, all numbers normalized w.r.t. fully connected TSV-based NoC.

Fig. 8. MTTT distribution of TSV-based VLs for RAD benchmark.

Fig. 9. Performance degradation with NFIC misalignment for 18 Gbps VL for CAN; all numbers normalized w.r.t. fully connected NFIC-based NoC with no link misalignment.

fail in the same amount of time. LB partially connected NoC have similar MTTT distribution to that of a PO fully connected NoC as discussed in Section 5.2.

However, there is one issue that affects the performance of NFIC-based VLs. This issue is misalignment of coils. But this does not pose severe communication failure as link communication can still be maintained albeit with a higher energy overhead. Figure 9 shows the effect of misalignment on NFIC-based NoCs; as an example, here, we have chosen to demonstrate the effects of misalignment for 18 Gbps NFIC-based VL. All the EDP numbers have been normalized with respect to a fully connected NFIC-based NoC of same data rate with no misalignment. As we can see here, with increase in misalignment, the EDP degrades as the energy consumption of NFIC VLs increases with misalignment. This shows that even with 40% misalignment the performance penalty is about 2% for fully connected NoCs and 4% for partially connected NoCs, which is insignificant.

5.4 Summary of Design Rules

From the above analysis and discussion, we can summarize the following design rules and associated trade-offs when designing 3D NoCs with NFIC- and TSV-based VLs.

1. When NFIC-based VLs have larger area overhead compared to TSV-based counterparts of same data rates, then a partially connected NFIC-based 3D NoC is viable. In this scenario, if the energy consumption is comparable between the two types of VLs, the partially connected NFIC-based NoC has equivalent performance to that of a fully connected TSV-based NoC.

2. When NFIC-based VLs have smaller area overhead compared to TSV-based VLs of same data rates, then a partially connected TSV-based 3D NoC is considered.
 - a. However, when the energy consumption of such NFIC VLs are not much higher than that of the TSV-based VLs, the fully connected NFIC-based NoC performs very close to a partially connected TSV-based NoC with same area overhead. The lack of EM-related workload-induced stress enabled by properly designed inductor coils make the NFIC-based architecture a more robust solution.
 - b. If the energy consumption of the NFIC-based VL is significantly higher than the TSV counterpart, then the partially connected TSV-based NoC significantly outperforms its fully connected NFIC-based counterpart with similar area overhead.
3. Physical characteristics of the VLs also dictate the optimization objective of the NoC design. The reliability concern of TSV-based VLs due to EM motivates us to undertake load-balancing-based optimization rather than optimizing only performance. This alleviates compromising the MTTF while designing a partially connected TSV-based NoC. However, NFIC-based VLs do not suffer from EM. Hence, we can optimize the performance exclusively.

6 CONCLUSIONS

VLs in 3D NoC affect the performance and reliability of the overall system. Both TSV- and NFIC-based links have specific advantages and limitations. Due to EM, TSV links suffer from workload-induced stress. However, NFIC links suffer from misalignment of inductor coils. However, workload-induced stress in TSV is much more severe than the misalignment problem in NFIC. Hence, we should balance the workload among the VLs in a TSV-based NoC so that the probability of EM-induced failure decreases. However, we need not address uneven link utilization in NFIC-based designs and we can focus on optimizing the overall performance irrespective of the load balancing. By adopting suitable performance optimization (load balancing vs. sole performance optimization), we can mitigate the performance degradation in partially connected NoCs and at the same time reduce the area overhead. With the adoption of appropriate optimization metrics based on the physical layer characteristics, the partially connected NoCs reduce area overhead by at least 15% while only sacrificing 1%–8% performance. Thus, our proposed optimization method enables the design of high performance and reliable partially connected NoC using both TSV- and NFIC-based links.

REFERENCES

- [1] D. Ditzel, T. Kuroda, and S. Lee. 2014. Low-cost 3D chip stacking with ThruChip wireless connections. In *Proceedings of the IEEE Hot Chips Symposium*. DOI: <https://doi.org/10.1109/HOTCHIPS.2014.7478813>
- [2] T. Kuroda. 2014. 3D Integration, power delivery, and contactless interconnect by near field coupling. In *Proceedings of the IEEE International Solid-state Circuits Conference*.
- [3] T. Frank et al. 2011. Resistance increase due to electromigration induced depletion under TSV. In *Proceedings of the International Reliability Physics Symposium*. 3F.4.1–3F.4.6. DOI: <https://doi.org/10.1109/IRPS.2011.5784499>
- [4] P. M. Yaghini, A. Eghbal, S. S. Yazdi, N. Bagherzadeh, and M. M. Green. 2016. Capacitive and inductive TSV-to-TSV resilient approaches for 3D ICs. *IEEE Trans. Comput.* 65 (2016), 693–705. DOI: <https://doi.org/10.1109/TC.2015.2498550>
- [5] Y. R. Lu, Y. Lee, W. J. McMahon, and T. Fung. 2006. Robust inductor design for RF circuits. In *Proceedings of the IEEE Custom Integrated Circuits Conference*. 571–574. DOI: <https://doi.org/10.1109/CICC.2006.320993>
- [6] D. Dutoit et al. 2013. A 0.9 pJ/bit, 12.8 GByte/s WideIO memory interface in a 3D-IC NoC-based MPSoC. In *Proceedings of the Symposium on VLSI Technology*. C22–C23.
- [7] S. Das, J. R. Doppa, D. H. Kim, P. P. Pande, and K. Chakrabarty. 2015. Optimizing 3D NoC Ddesign for energy efficiency: A machine learning approach. In *Proceedings of the IEEE/ACM International Conference on Computer-aided Design (ICCAD'15)*. IEEE Press, 705–712. DOI: <https://doi.org/10.1109/ICCAD.2015.7372639>

- [8] D. Lee, S. Das, and P. P. Pande. 2019. Analyzing power-thermal-performance trade-offs in a high-performance 3D NoC architecture. *Integration* 65 (2019), 282–292. DOI : <https://doi.org/10.1016/j.vlsi.2017.12.002>
- [9] D. Lee, S. Das, and P. P. Pande. 2017. Performance-thermal trade-offs for a VFI-enabled 3D NoC architecture. In *Proceedings of the 18th International Symposium on Quality Electronic Design (ISQED'17)*. 271–276. DOI : <https://doi.org/10.1109/ISQED.2017.7918327>
- [10] M. Pathak, J. Pak, D. Z. Pan, and S. K. Lim. 2011. Electromigration modeling, and full-chip reliability analysis for BEOL interconnect in TSV-based 3D ICs. In *Proceedings of the IEEE/ACM International Conference on Computer-aided Design (ICCAD'11)*. 555–562. DOI : <https://doi.org/10.1109/ICCAD.2011.6105385>
- [11] L. Doyen, E. Petitprez, P. Waltz, X. Federspiel, L. Arnaud, and Y. Wouters. 2008. Extensive analysis of resistance evolution due to electromigration induced degradation. *J. Appl. Phys.* 104 (2008), 123521. DOI : <https://doi.org/10.1063/1.3043798>
- [12] S. Das, J. R. Doppa, P. P. Pande, and K. Chakrabarty. 2016. Reliability, and performance trade-offs for 3D NoC-enabled multicore chips. In *Proceedings of the Design, Automation & Test in Europe Conference & Exhibition (DATE'16)*. 1429–1432.
- [13] S. Das, J. R. Doppa, P. P. Pande, and K. Chakrabarty. 2017b. Robust TSV-based 3D NoC design to counteract electromigration, and crosstalk noise. In *Proceedings of the Design, Automation & Test in Europe Conference & Exhibition (DATE'17)*. 1366–1371. DOI : <https://doi.org/10.23919/DATE.2017.7927205>
- [14] S. Das, J. R. Doppa, P. P. Pande, and K. Chakrabarty. 2017a. Design-space exploration and optimization of an energy-efficient and reliable 3-D small-world network-on-chip. *IEEE Trans. Comput.-aided Des. Integr. Circ. Syst.* 36 (2017), 719–732. DOI : <https://doi.org/10.1109/TCAD.2016.2604288>
- [15] Q. Gu, Z. Xu, J. Ko, and M. F. Chang. 2007. Two 10Gb/s/pin low-power interconnect methods for 3D ICs. In *Proceedings of the IEEE International Solid-state Circuits Conference*. 448–614. DOI : <https://doi.org/10.1109/ISSCC.2007.373487>
- [16] Y. Take, H. Matsutani, D. Sasaki, M. Koibuchi, T. Kuroda, and H. Amano. 2014. 3D NoC with inductive-coupling links for building-block SiPs. *IEEE Trans. Comput.* 63 (2014), 748–763. DOI : <https://doi.org/10.1109/TC.2012.249>
- [17] D. H. Kim and S. K. Lim. 2015. Physical design and CAD tools for 3-D integrated circuits: Challenges and opportunities. *IEEE Des. Test* 32 (2015), 8–22. DOI : <https://doi.org/10.1109/MDAT.2015.2440317>
- [18] T. Kuroda. 2010. Thruchip interface (TCI) for 3D integration of low-power system. In *Proceedings of the International Electron Devices Meeting*. 17.1.1. DOI : <https://doi.org/10.1109/IEDM.2010.5703378>
- [19] S. Gopal, S. Das, P. Agarwal, S. N. Ali, D. Heo, and P. P. Pande. 2018. High-performance and small-form factor near-field inductive coupling for 3-D NoC. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* 26 (2018), 2921–2934. DOI : <https://doi.org/10.1109/TVLSI.2018.2865704>
- [20] Y. Take, N. Miura, and T. Kuroda. 2011. A 30 Gb/s/Link 2.2 Tb/s/mm² Inductively-coupled injection-locking CDR for high-speed DRAM interface. *IEEE J. Solid-state Circ.* 46 (2011), 2552–2559. DOI : <https://doi.org/10.1109/JSSC.2011.2164023>
- [21] S. Das, S. Gopal, D. Heo, and P. P. Pande. 2017. Energy-efficient and robust 3D NoCs with contactless vertical links. In *Proceedings of the IEEE/ACM International Conference on Computer-aided Design (ICCAD'17)*. 1–6. DOI : <https://doi.org/10.1109/ICCAD.2017.8501894>
- [22] S. Gopal, S. Das, D. Heo, and P. P. Pande. 2017. Energy and area-efficient near field inductive coupling: A case study on 3D NoC. In *Proceedings of the 11th IEEE/ACM International Symposium on Networks-on-Chip (NOCS'17)*. Association for Computing Machinery, New York, NY. 1–8. DOI : <https://doi.org/10.1145/3130218.3130224>
- [23] H. Matsutani, M. Koibuchi, I. Fujiwara, T. Kagami, Y. Take, T. Kuroda, P. Bogdan, R. Marculescu, and H. Amano. 2014. Low-latency wireless 3D NoCs via randomized shortcut chips. In *Proceedings of the Design, Automation & Test in Europe Conference & Exhibition (DATE'14)*. 1–6. DOI : <https://doi.org/10.7873/DATE.2014.286>
- [24] M. Koibuchi, L. Leong, T. Totoki, N. Niwa, H. Matsutani, H. Amano, and H. Casanova. 2019. Sparse 3-D NoCs with inductive coupling. In *Proceedings of the 56th Annual Design Automation Conference (DAC'19)*. Association for Computing Machinery, New York, NY. 1–6. DOI : <https://doi.org/10.1145/3316781.3317913>
- [25] F. Dubois, A. Sheibanyrad, F. Pétrot, and M. Bahmani. 2013. Elevator-First: A deadlock-free distributed routing algorithm for vertically partially connected 3D-NoCs. *IEEE Trans. Comput.* 62 (2013), 609–615. DOI : <https://doi.org/10.1109/TC.2011.239>
- [26] S. Das, D. Lee, D. H. Kim, and P. P. Pande. 2015. Small-world network enabled energy-efficient and robust 3D NoC architectures. In *Proceedings of the 25th Great Lakes Symposium on VLSI (GLSVLSI'15)*. Association for Computing Machinery, New York, NY. 133–138. DOI : <https://doi.org/10.1145/2742060.2742085>
- [27] P. Bogdan. 2015. Mathematical modeling and control of multifractal workloads for data-center-on-a-chip optimization. In *Proceedings of the 9th International Symposium on Networks-on-Chip (NOCS'15)*. Association for Computing Machinery, New York, NY. 1–8. DOI : <https://doi.org/10.1145/2786572.2786592>
- [28] Y. Xue and P. Bogdan. 2015. User cooperation network coding approach for NoC performance improvement. In *Proceedings of the 9th International Symposium on Networks-on-Chip (NOCS'15)*. Association for Computing Machinery, New York, NY. 1–8. DOI : <https://doi.org/10.1145/2786572.2786575>

- [29] K. Duraisamy, Y. Xue, P. Bogdan, and P. P. Pande. 2017. Multicast-aware high-performance wireless network-on-chip architectures. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* 25 (2017), 1126–1139. DOI : <https://doi.org/10.1109/TVLSI.2016.2612647>
- [30] M. Saito, Y. Sugimori, Y. Kohama, Y. Yoshida, N. Miura, H. Ishikuro, T. Sakurai, and T. Kuroda. 2010. 2 Gb/s 15 pJ/b/chip inductive-coupling programmable bus for N, and flash memory stacking. *IEEE J. Solid-state Circ.* 45 (2010), 134–141. DOI : <https://doi.org/10.1109/JSSC.2009.2034431>
- [31] M. Saito, Y. Sugimori, Y. Kohama, Y. Yoshida, N. Miura, H. Ishikuro, and T. Kuroda. 2009. 47% power reduction and 91% area reduction in inductive-coupling programmable bus for N, and flash memory stacking. In *Proceedings of the IEEE Custom Integrated Circuits Conference*. 449–452. DOI : <https://doi.org/10.1109/CICC.2009.5280819>
- [32] C. Cassidy et al. 2012. Through silicon via reliability. *IEEE Trans. Dev. Mater. Reliab.* 12 (2012), 285–295. DOI : <https://doi.org/10.1109/TDMR.2012.2189212>
- [33] D. H. Jung et al. 2015. Modeling and analysis of defects in through silicon via channel for non-invasive fault isolation. In *Proceedings of the International 3D Systems Integration Conference (3DIC'15)*. TS8.29.1–TS8.29.4. DOI : <https://doi.org/10.1109/3DIC.2015.7334599>
- [34] G. V. D. Plas et al. 2011. Design issues and considerations for low-cost 3-D TSV IC technology. *IEEE J. Solid-state Circ.* 46 (2011), 293–307. DOI : <https://doi.org/10.1109/JSSC.2010.2074070>
- [35] U. Y. Ogras and R. Marculescu. 2006. “It’s a small world after all”: NoC performance optimization via long-range link insertion. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* 14 (2006), 693–706. DOI : <https://doi.org/10.1109/TVLSI.2006.878263>
- [36] B. K. Joardar, R. G. Kim, J. R. Doppa, P. P. Pande, D. Marculescu, and R. Marculescu. 2019. Learning-based application-agnostic 3D NoC design for heterogeneous manycore systems. *IEEE Trans. Comput.* 68, 6 (June 2019), 852–866. DOI : <https://doi.org/10.1109/TC.2018.2889053>
- [37] N. Binkert et al. 2011. The Gem5 simulator. *SIGARCH Comput. Archit. News* 39, 2 (Aug. 2011), 1–7. DOI : <https://doi.org/10.1145/2024716.2024718>
- [38] B. S. Feero and P. P. Pande. 2009. Networks-on-chip in a three-dimensional environment: A performance evaluation. *IEEE Trans. Comput.* 58 (2009), 32–45. DOI : <https://doi.org/10.1109/TC.2008.142>
- [39] R. G. Kim, W. Choi, Z. Chen, P. P. Pande, D. Marculescu, and R. Marculescu. 2016. Wireless NoC and dynamic VFI codesign: Energy efficiency without performance penalty. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* 24 (2016), 2488–2501. DOI : <https://doi.org/10.1109/TVLSI.2015.2512611>
- [40] S. Bandyopadhyay, S. Saha, U. Maulik, and K. Deb. 2008. A simulated annealing-based multiobjective optimization algorithm: AMOSA. *IEEE Trans. Evolut. Comput.* 12 (2008), 269–283. DOI : <https://doi.org/10.1109/TEVC.2007.900837>
- [41] N. Agarwal, T. Krishna, L. Peh, and N. Jha. 2009. GARNET: A detailed on-chip network model inside a full-system simulator. In *Proceedings of the IEEE International Symposium on Performance Analysis of Systems and Software*. 33–42. DOI : <https://doi.org/10.1109/ISPASS.2009.4919636>
- [42] O. Lysne, T. Skeie, S. Reinemo, and I. Theiss. 2006. Layered routing in irregular networks. *IEEE Trans. Parallel Distrib. Syst.* 17 (2006), 51–65. DOI : <https://doi.org/10.1109/TPDS.2006.12>
- [43] S. C. Woo, M. Ohara, E. Torrie, J. P. Singh, and A. Gupta. 1995. The SPLASH-2 programs: Characterization and methodological considerations. In *Proceedings of the 22nd Annual International Symposium on Computer Architecture*. 24–36. DOI : <https://doi.org/10.1109/ISCA.1995.524546>
- [44] Christian Bienia. 2011. Benchmarking modern multiprocessors. Ph.D. Dissertation. Princeton University, NJ.
- [45] Y. Liu, W. Luk, and D. Friedman. 2012. A compact low-power 3D I/O in 45nm CMOS. In *Proceedings of the IEEE International Solid-state Circuits Conference*. 142–144. DOI : <https://doi.org/10.1109/ISSCC.2012.6176900>

Received January 2020; revised April 2020; accepted April 2020