International Journal of Computer Applications (0975 - 8887)
Volume 121 - No. 6, July 2015

A Study of Optimization Techniques for 3D Networks-on-Chip Architectures for Low Power and High Performance Applications

Michael Opoku Agyeman
Department of Computer Science and Engineering
The Chinese University of Hong Kong

ABSTRACT

Three dimensional Networks-on-Chip (3D NoCs) have attracted a growing interest to solve on-chip communication demands of future multi-core embedded systems. However, 3D NoCs have not been completely accepted into the mainstream due to issues such as the high cost and complexity of manufacturing 3D vertical wires, larger memory, area and power consumption of 3D NoC components than that of conventional 2D NoC. This paper presents a brief about 3D NoCs optimization techniques with focus on modeling and evaluation of alternate NoC topologies, routing algorithms and mapping techniques to achieve optimized area, power and performance parameters (latency and throughput). Particularly, we investigate novel 3D NoC router architectures and their possible combinations which aim at achieving lower area and power consumption of on-chip communication components with a minimal performance trade-off.

Keywords

Network-on-Chip, System-on-Chip, 3D Integration, Low Power

1. INTRODUCTION

With the promising ability to provide high computational power on the go, embedded systems are widely applied in many everyday life activities including portable electronics devices such as smart phones, home electronics, medical technology, industrial automation, multimedia, telecommunication and avionics systems. As an effort to provide high performance computing within the current time-to-market constraints, Systems-on-Chip (SoC) emerged as an embedded systems technology where several components such as processors, power management circuits, memories and image processors are placed on a single integrated circuit (IC). Moreover, with the continuous demand for high performance, high density, portable and less power hungry electronic devices combined with the endless growth in the power consumption and diminishing performance of single processor architectures, multi-core systems have emerged as an inevitable SoC solution.

A dominant issue with next generation of multi-core IC design arises from the non-scalable wire delays and power consumption [1]. These issues, such as the maximum number of cores per shared bus, reliability and efficient arbitration for accessing shared bus, have attracted a lot of research interest over the past few years [1]. Network-on-Chip (NoC) has been proposed as a more promising solution and has gained the attention of many researchers in this field of study [2]. NoCs adopt links, switches and packet or circuit switching to handle data communication in a multi-core system. The main idea of NoCs is to reduce the disadvantages of conventional SoC communication architectures such as high latency in buses and the numerous long interconnects in *point-to-point* (P2P) communication.

Three dimensional integrated circuit (3D IC) design is another emerging technology, where several 2D silicon layers are stacked vertically. 2D ICs have larger footprint than comparative 3D implementations. The shorter and more efficient vertical wires provided by the 3D design to replace the long horizontal 2D wires presents lower power consumption and interconnect delays. Also besides the communication performance advantages, 3D IC allows heterogeneity as different set of devices and disparate technologies can be integrated on different silicon layers. A popular choice for 3D integration is the Wafer-to-wafer bonding technology, where the vertical interconnects are implemented using Through Silicon Vias (TSVs) [3]. However, due to insufficient 3D design and synthesis tools as well as the limitations of TSVs coupled with the extra manufacturing processes such as wafer thinning 3D IC technology is yet to be generally accepted by the industry.

3D NoCs have been extensively investigated as a promising technology to extend the beneficial attributes provided by the design of 3D multi-core chips. NoCs are scalable, provide configurable parallelism and control over interlayer vertical interconnects (TSVs) [4]. Combining NoCs and 3D integration technology (3D NoCs) introduces new opportunities and design challenges. Example, the challenging task of designing power-efficient NoCs for 3D multi-core systems that meet both the performance requirements of applications and technology constraints. 3D NoC architecture has received some research attention [5–8]. However, little effort has been made to reduce the cost of 3D-links in terms of reducing power consumption and increasing performance by adopting hybrid interconnection paradigm that use a mixture of different router architectures. This requires new design methods for 3D NoCs considering the 3D constraints on interlayer communication, the number of supported TSVs and router placement in the layers of the 3D structure.

2. MOTIVATION AND SCOPE OF STUDY

2.1 Motivation

Several investigations have been made to address performance and technological constraints associated with 3D NoC topology design. Core assignment and floorplanning of different layers of the 3D chip need to take several performance and technological constraints such as thermal issues into consideration [9]. Also in many designs, the layer assignment of cores is dictated by technology constraints, e.g., having logic and memory on different layers [10].

Different cores in a typical multi-core chip can have different bandwidth and latency requirements. For efficient on-chip

communication, it is essential that these parameters are taken into consideration at early design stage. Placing switches closer to cores with high bandwidth requirements provides shorter links for high traffic loads with better energy and latency values compared to having equal distances between all cores and switches. Also TSVs have better delay characteristics than the local horizontal links [8]. With this in mind, cores in a 3D architecture can be partitioned in various ways and switches assigned accordingly. Partitioning cores and placement of cores and switches have been investigated with considerations to different technological constraints. For example communication graphs, local partitioning graphs and scaled partitioning graphs [10] have been used for core to switch connectivity in order to increase performance in application-specific multi-core architectures. Algorithms for path computation, establishing number of switches and switch position computation for application-specific 3D-NoCs synthesis that takes technology and performance constraints into account are also presented in [10].

Different router architectures have different power and performance impact on the on-chip communication resources. A study of 3D NoC with distributed 2D and symmetric 3D router per 3D layer is presented in [11]. The investigation was focused on distributing the routers and studying performance and area parameters under two separate 3D NoC topologies (homogeneous 3D mesh and 3D torus). The study does not consider different router designs, an appropriate mix of different topologies and hybrid NoC interconnect paradigms. Similarly, in [10, 11], the research adopts router architectures from [12] which implements symmetric 3D NoC router. The addition of two extra ports to a conventional 2D router is claimed to cause a rise of 102.8% and 125.3% in the crossbar area and power consumption respectively [8, 13]. Less power hungry and high performance router designs such as 3D NoC-bus hybrid router will be investigated in the scope of this study.

One of the significant drawbacks to 3D IC design is the constraint on the number of TSVs per IC layer. The number of TSVs is restricted due to the large area overhead of the via pads required for interfacing the TSVs in each active layer. Also due to wafer misalignment issues with current TSV manufacturing technology, TSV fabrication has low yield [14]. A straight forward solution to improve the yield is to use TSV redundancy [15]. An alternative option will be to use larger bonding pads to reduce the probability of inevitable shifts of the via pads [16]. Either solution introduces more resources which increases the area and power consumption of the NoC. Also, large pads increase the routing complexity of the active layers. Therefore, there is the need for novel 3D NoC architectures that focus on reducing the area overhead of via pads in the active layers without sacrificing the performance of 3D NoCs.

Besides the difference in total power consumption and latency from one topology to the other, the architecture of different NoC components also has different impact on the total floorplan area, power consumption and the latency [8], such as number of central arbiters, crossbar connections per switch, number of ports per 3D switch, routing algorithms of the 3D and 2D routers, etc. An appropriate mix of different NoC components will be studied. Combining different components such as 2D and 3D routers to work efficiently in one 3D NoC design requires intelligent deadlock free, live-lock free and efficient routing algorithms. New static, oblivious and adaptive routing algorithms for such architectures as well as their implementation in hardware needs to be investigated to ensure high communication performance and practicality.

The main objective of this study is to explore and present generic solutions with 3D NoC technology and performance constraints which provide an optimized architecture that resolves the issue of 3D router area and power overheads as well as expensive TSV manufacturing cost. Issues to be considered include: latency, symmetry, throughput, power consumption and packet energy. The investigation is expected to present critical analysis of heterogeneous schemes with low power-high performance architectures by investigating new combinations of 2D and 3D routers in 3D NoCs.

Efficient mapping algorithms and performance trade-off associated with exploring new partitions is an open NoC research area that needs intensive investigation. For example, cores with high bandwidth demands can be grouped with cores with low bandwidth requirements in the same partition to balance the traffic load in the NoC while reducing hotspots and hop-count. Moreover, combining 2D and 3D routers in NoCs introduces new application mapping challenges. Questions such as, which mapping approach can generate an optimal combination of 2D and 3D routers without sacrificing the performance of homogeneous 3D NoCs must be investigated. Also the issue of how can applications be mapped to improve the performance of such architectures while reducing the power consumption for any giving routing algorithm has to be addressed.

2.2 Scope

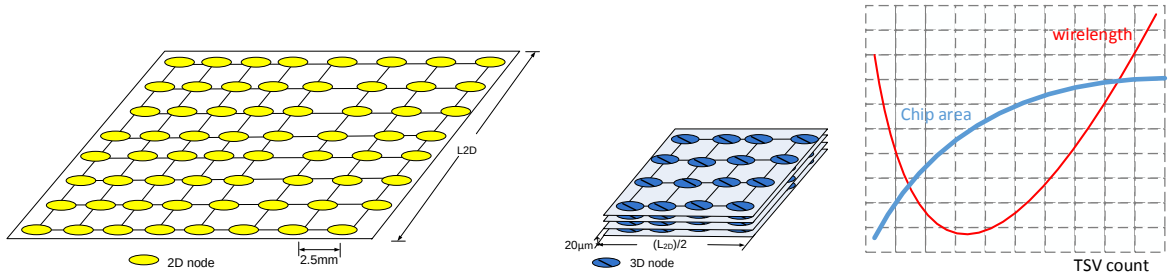
In this paper, the emphasis of the work is to improve the area, power and manufacturing cost of 3D NoCs without sacrificing the performance of the network. Alternative NoC architectures are modeled and evaluated by employing heterogeneous 3D NoC architectures and intelligent routing strategies for future multi-core design. Specifically, NoC components are investigated and exploited to offer various means of improving the performance and cost constraints of on-chip-communication. In relation to this problem statement, this paper is based on existing and current work on 2D and 3D NoC architectures design with more focus on the combination of low power and high throughput NoC router architectures with minimal but efficient vertical interconnects for 3D NoC structures. The aforementioned performance and cost constraints for NoCs have recently attracted significant interests from researchers worldwide. The objectives can be summarized as follows:

(1) To study novel hybrid interconnection paradigm that incorporates an appropriate mix of different router architectures in terms of area, latency and power efficiency with considerations of minimizing number of 3D vertical interconnects and investigating the feasibility of applying new hybrid interconnection paradigm to multi-core systems.

(2) To investigate intelligent routing algorithms that ensure efficient intra- and interlayer data communication.

(3) To explore generic approaches that automatically generate high performance heterogeneous 3D NoC architectures for a given application. Particularly, the aim of this objective is to explore solutions to that improve the power efficiency of 3D NoCs by exploiting utilization of various NoC components in a given application.

(4) To study efficient application mapping techniques in 3D NoCs with limited number of vertical links.



(a) 8×8 2D Mesh

(b) $4 \times 4 \times 4$ 3D Mesh

(c) Variation of total wirelength and chip area with TSV count [19]

Fig. 1. Comparison between 2D and 3D mesh.

3. 3D NOC ARCHITECTURES

To optimize the long interconnect wires and larger footprint of SoC design on the traditional 2D IC, 3D integration technology has been proposed to stack vertically several of 2D silicon layers. As shown in Figure 1, the layers in the 3D IC provide shorter interlayer wires and more resource connectivity with reduced hop-count compared to the long interconnect wires in a 2D SoC implementation with equivalent number of resources [17]. Moreover, 3D technology allows heterogeneous integration of technologies and design disciplines such as RF, digital, analog, etc. on a single chip [18]. To meet the constraints of SoC design in the third dimension, Network-on-Chip (NoC) has been proposed as a promising solution [4]. NoCs are scalable and helpful in providing configurable parallelism and control over the interlayer vertical interconnects such as Through Silicon Vias (TSVs) and intra-layer wires [2,4]. Combining NoCs and 3D integration technology (a.k.a 3D NoCs) introduces new opportunities and design challenges [13]. One of the main design challenges is TSV manufacturing which is an expensive and complicated process [20]. Additionally, TSV manufacturing processes have high defect rates which result in poor yields [21]. Also TSV pads required for bonding consume a considerable amount of area in each layer of the 3D chip [22]. However, most of the work on 3D NoCs in literature assumes full vertical connectivity. This may not be optimal for many applications, as 3D NoC router has a larger area footprint and more energy consumption than a generic 2D router [5, 8]. So it is crucial to use 3D routers with vertical interconnections as efficiently as possible, which is one of the objectives of this study.

3D circuit opens a huge research space for multi-core design and CAD algorithms, specifically in the NoCs area. New algorithms must be provided in order to efficiently handle 3D placement of circuit elements and to take full benefit of technology while considering new issues caused by the 3D integration [23]. Consequently, we investigate alternative architectures for low power and area efficient 3D NoC implementation by employing 2D routers and single hop 3D NoC-bus hybrid routers for interlayer communications. Reducing the number of 3D routers may result in higher delay in conveying packets to their destinations. Hence, the second focus is to investigate trade-off analysis of such architectures in terms of average packet latency, energy consumption, and area overhead. Thirdly, deadlock free shortest path deterministic algorithm for efficiently routing packets in such 3D NoCs is investigated.

3.1 3D Integration Technology

Compelled by increasing thermal, power, financial to performance trade-off beyond 28nm, the industry has resorted to thinned and stacked 3D ICs as an alternative IC manufacturing technology [19]. 3D IC manufacturing has evolved from early wire-bond to flip-chip and recently Through-Silicon Via (TSV). With TSVs, digital and analog functional blocks can be arranged across multiple dies at a much fined granularity introducing shorter wire lengths with lower power consumption. However, according to ITRS [24], the size of TSVs diameter is expected to range from 1.5um to 1.0um between 2009 and 2015 while for the same period the area of a 4-transistor logic gate is expected to range from 0.82um to 0.20um. Thus, over 100% increase is expected in the area ratio of TSVs and logic gates. An even larger ratio is expected if the keep-out zone around the TSVs is considered, which is required to minimize manufacturing issues such as stress and lithography [19]. TSVs are vertical links in the third dimension that completely pass through a silicon die mainly to provide electrical connectivity between devices in two different dies in a 3D IC. The major TSV fabrication technologies available are Via-first and Via-last. In via-first, TSVs are fabricated before BEOL (back-end-of-line) metallization while in via-last, TSVs are created after BEOL or bonding. Via-last TSV has a wider diameter (usually between 10-50um) compared to via-first (usually between 1-10um) [20, 25]. We investigate relevant efforts in literature that have made advances to reduce the area impact of TSV fabrication (particularly, Via-last TSVs) on 3D Chips.

3.2 Analysis of Through Silicon Vias

TSV is generally accepted as the most viable solution for 3D IC design. However, several concerns such as TSV count and placement have drawn a lot of attention as they affect the quality and yield of 3D ICs. In a typical 3D chip, different TSV types, such as control, power distribution and data transmission TSVs may be employed which each of them has a different bandwidth. It is reported that the total wire-length (to a certain degree) reduces as the TSV count per 3D IC increases [26]. On the other hand, due to the large size of the TSV bonding pads, the total wire-length increases dramatically after the TSV count exceeds a critical value. Also as shown in Figure 1(c), chip area increases as the TSV count increases [27]. However, the number of TSVs in a 3D IC can be reduced if the TSVs are efficiently distributed among the on-chip resources. Moreover, the total cost of 3D ICs increases as the TSV count increases [28].

Consequently, impact of TSV count on 3D IC design quality and yield opens several challenges in both 3D NoC and IC design technologies. This calls for investigating heterogeneous or irregular topologies with focus on minimizing the number of interlayer links across 3D layers, while presenting enhanced performance and cost parameters [29].

4. OPTIMIZED HETEROGENEOUS 3D NOC ARCHITECTURES

Generally, 2D NoCs suffer from long interconnect delays and large floorplan area as the number of cores increases. Standard homogeneous 3D NoC architectures have been proposed as an effective solution to the on-chip communication delays by stacking several 2D layers while employing vertical links at each node for interlayer communication [13].

The topology of a network affects its performance and power consumption. Feero et al. [17] evaluated strategies for comparing the performance characteristics of NoC architectures under various 3D homogeneous topologies. Also, Pavlidis et al. [5] presented an analytical model for zero-load latency that considers the effect of topology on 3D NoC. Their work emphasizes on regular, homogeneous NoC topologies and their performance metrics.

Future SoCs will have heterogeneous cores and components which will require optimal heterogeneous NoC topologies [25]. By redistributing the router buffers and the interconnect link widths, Mishra et al. proposed HeteroNoC, a heterogeneous architecture for 2D NoCs [30]. With this approach some routers ports have more virtual channels with wider interconnect links than others. This approach involves repacketization at different NoC regions to enable routing along different link widths. Virtual channel implementation consumes a lot of buffer and suffers from starvation for some NoC applications [31]. Besides, such routers suffer from large area and power consumption [32]. A study of the area and energy benefits of combining 2D and 3D routers in 3D mesh and torus topologies is presented in [11]. In their work, the placement of 3D routers does not consider the traffic load of the selected tile. These architectures will not perform well under high traffic conditions in different traffic patterns. Secondly, they do not consider scalability (network size), and practical benchmarking in 3D NoC architectures. Xu et al. [33] performed an evaluation of the impact of reducing the number of TSVs to half and quarter on the performance of 3D NoCs. Their proposed architectures, quarter/lo and half/lo (quarter/hi and half/hi), aim at generating heterogeneous 3D NoC with 2D routers placed as close to (far from) 3D routers as possible in each layer. These architectures suffer from uneven distribution of 3D routers and unpredictable delays for different applications. Both [11] and [33] analysed their 3D NoC architectures with a mix of 2D routers and symmetric 3D routers in terms of total performance. However with increase in number of ports, the crossbar power consumption and occupied area increases significantly when symmetric 3D routers are used to replace conventional 2D routers [8].

Li et al. proposed to replace the large 7 port symmetric 3D routers with 6 port NoC-bus hybrid 3D routers introducing the Hybrid 3D NoC-bus mesh architecture [34]. Here, dynamic Time-Division Multiple Access (dTDMA) bus is used for single hop interlayer communication while NoC is used for multi-hop intra-layer communication. This architecture requires an addition of a central arbiter per each vertical pillar to allow seamless integration of the Bus and NoC interface. Besides the additional resource required, the 6 port 3D router still has a large crossbar and energy consumption.

Liu et al. [22] used partition islands of routers to constitute regions for sharing the same TSV pad for interlayer communication controlled by serialization logic. However, serialization along the TSV bundle causes the average packet delay to increase exponentially as the number of routers per TSV bundle increases. Moreover, due to serialization logic, the TSV pads have no direct connection to the processing cores, which is a waste of chip area.

The goal of this section is to investigate different heterogeneous 3D NoC architectures, that employ small footprint 5 port 2D routers to provide intra-layer hop by hop communication and 6 port 3D NoC-bus hybrid routers to provide single hop interlayer communication, and analyse their average packet latency (performance saturation rates), energy consumption, area overhead and scalability (network size) under different synthetic and real-world traffic patterns.

Hence, more efficient architectures should be designed. Considering the manufacturing cost of 3D NoCs, [35–37] presents different area efficient and low power heterogeneous 3D NoC architectures, which combine both the power and performance benefits of 2D routers and 3D NoC-bus hybrid router architectures in 3D NoC architectures. Experimental results show a negligible penalty (less than 5%) in average packet latency of the heterogeneous 3D NoC architectures compared to typical homogeneous 3D NoCs, while the heterogeneity provides power and area efficiency of up to 61% and 20%, respectively.

5. HIGH PERFORMANCE ROUTING TECHNIQUES FOR HETEROGENEOUS 3D NOCS

Packets in heterogeneous 3D NoC architectures travel along longer paths due to the limited access to other layers. Also, routing of traffic along different layers for heterogeneous 3D NoC architectures can be challenging due to the limited number of vertical links. Kim et al. [8] proposed DimDe, a dimension decomposed router architecture to optimize the area and power consumption of 3D NoC switches and it is limited to homogeneous NoCs. This router would not be suitable for most real-world applications in 3D NoCs as they are more likely to be heterogeneous with mixed memory and logic layers [3, 13, 25]. Liu et al. [22] have demonstrated that the average utilization of the TSVs is significantly minimal for homogeneous 3D mesh under non-uniform low traffic conditions. Hence, we focus on heterogeneous 3D NoCs with limited number of TSVs.

Xu et al. [33] performed an evaluation of the impact of reducing the number of TSVs to half and quarter on the performance of 3D NoCs. Here, packets are routed along the x dimension and as long as they meet a pillar, the packets will be routed along the y dimension to the pillar and finally to the destination layer. This routing scheme suffers from unpredictable contention issues for non-minimal hop heterogeneous architectures.

An XYZ static routing algorithm for 3D mesh with 2D and 3D routers is presented in [11]. In their work, the routers have no knowledge of the traffic load of the selected pillar and packets will always be routed along the same shortest path. These algorithms will not perform well under high traffic conditions. Routing packets along pillars with minimal paths distorts the balance of the network traffic, even under uniform traffic with evenly distributed 3D routers for heterogeneous 3D NoCs. Moreover, other paths that can be used to improve the NoC performance are underutilized while contention along some pillars increases exponentially. Ying et al. [38] presented routing techniques for 3D NoCs with reduced vertical links where interlayer packets are either

sent along 3D routers with shortest distance to the source node or destination node. Also for deadlock freedom in adaptive routing, they presented a quadrant based routing (turn model) algorithm to send packets along 3D routers in North-East (South-East, North-West or South-West) direction with shortest path between source and destination nodes. However, shortest distance to source or destination may increase contention as packets can be sent along long paths between the source and destination nodes. Also, quadrant based routing requires that a 3D router is placed in a quadrant that favours the allowed turn. However, various mapping and placement algorithms generates different architectures which may be impractical with the routing technique.

Rahmani et al. [39] proposed a fault tolerant adaptive routing algorithm for homogeneous 3D NoC with 3D Bus-hybrid routers. By removing router output buffers and adopting virtual channels, the adaptive router architecture avoids deadlocks and detects faulty vertical wires based on the status of a wait signal provided by the central arbiter. These algorithms will experience long packet delays under heterogeneous architectures as extra clock cycles and implementation logic will be wasted to identify non-existing vertical links even if they can be considered as faulty links. Also, extra delays will be introduced to reroute packets that were sent along these paths.

Most of the work on adaptive routing so far are designed for homogeneous 3D NoC architectures [39]. Adaptive routing algorithms for heterogeneous 3D mesh architectures will give better performance since the router architecture has knowledge of the location of the vertical pillars and will not even consider routing packets along non-existing pillars. Hence, we investigate efficient adaptive router for heterogeneous 3D NoC topologies with reduced number of vertical interconnects.

In [40, 41] an adaptive router that forwards packets towards the vertical link whose path provides the minimum delay as well as minimum Manhattan distance to the destination is proposed. Also, the work presents oblivious routing techniques that send packet along different equal cost vertical links in a predefined fashion. Furthermore [40] presents the hardware implementation details of area, power consumption and operating frequency of proposed router architectures. Experimental results show that their proposed architecture significantly improves the performance up to 75% by replacing 2D static routers with adaptive 2D routers in heterogeneous 3D NoCs, while keeping the maximum clock frequency, power and energy consumption of the adaptive router nearly at the same level as the static router.

6. DESIGN AUTOMATION IN HETEROGENEOUS 3D NOCS

The design practicality of 3D NoCs faces several challenges such as thermal issues, high power consumption and area of the conventional 3D router, high complexity and cost of vertical link implementation. Various 3D NoC topologies are presented and evaluated in [6,42] where homogeneous 3D routers are employed in each architecture. Pasricha [43] proposed a serialization technique for reducing the number of TSVs where link size of TSVs at selected nodes is reduced by a fraction. Thus if the number of TSVs exceeds a threshold, serialization is adopted to reduce the bandwidth of some TSVs. However, due to the reduced bandwidth of the TSVs and serialization logic, such architectures have high average packet latencies. Moreover, due to the higher overhead of serialization receiver and transmitter logic compared to the TSV reduction, such architectures have even higher power consumption compared to homogeneous 3D mesh. Based on the serialization methodology, Pasricha [44] proposed a 3D NoC synthesis framework by augmenting router and placement techniques proposed in [10], these routers have several local ports which have high power consumption due to the increased number of ports and high data rates across the crossbar.

Various buffer sizing algorithms have been presented based on probabilistic modelling [45–47]. Models presented in their work are based on queuing theory which restricts the input traffic injection distribution to adhere to a probabilistic pattern (Poisson distribution in this case). To enhance the performance of various applications in NoCs by efficiently redistributing the buffer spaces, it is crucial to accurately model congestion and hotspot regions in different applications. Kumar et al. [48] proposed a buffer-sizing algorithm for application-specific NoCs with multiple voltage-frequency islands. Here, the buffers are increased iteratively based on the NoC's behaviour under simulation. However, due to the iterative approach, the run-time of the simulation increases significantly as the number of nodes in the NoC increases. Xu et al. [49] presented an approach where heterogeneous architectures are generated by evenly distributing 3D and 2D routers in 3D NoCs. Here multiple 3D routers (redundant routers) are placed equidistant to the 2D routers. However, the generation of these architectures has no knowledge of the communication dynamics target traffic and assumes a uniform traffic distribution which is not the case in real world applications.

A systematic approach to generate efficient heterogeneous architectures that anticipate performance of target 3D NoC applications by placing 3D routers at highly utilized vertical links is presented in [50]. Moreover by exploiting average buffer utilization of each router, their method judiciously assigns more buffer resources to highly utilized channels in the routers while reducing that of lowly utilized ones without the use of virtual channel or the need for alteration of the routing algorithm. Experimental results in [50] show that the systematic algorithm generates optimized architectures with lower energy consumption and a significant reduction in packet delay compared to the hop-count based heterogeneous architectures. An efficient application mapping on heterogeneous 3D NoCs can be complex, partly due to the limited number of vertical links and also due to the fact that most applications cannot be completely characterized before design time. However, application mapping has a great impact on the performance, reliability and energy consumption of NoCs.

Several IP mapping algorithms have been proposed for 2D NoCs that focus on minimizing the overall communication power [51–53]. However, a simple extension of most of these existing mapping algorithms for 2D NoCs to 3D NoCs is not ideal as communication in the third dimension introduces a new set of NoC constraints which are not considered in the 2D designs. Moreover, such mapping algorithms will not be suitable for generation of 3D NoCs with heterogeneous router distributions as they do not take advantage of the power and performance benefits of the heterogeneity. Branch-and-Bound algorithm proposed by Hu et al. [52] can be extended to map application to heterogeneous 3D NoCs. However, this algorithm employs partial exhaustive search trees and has an extremely long run-time. Janidarmian et al. [54] proposed Onyx, a heuristic bandwidth-constrained mapping algorithm for tile-based NoCs. Here, nodes with higher communication data rates are mapped first. Tosun [55] presented CastNet, a heuristic algorithm for mapping tasks onto mesh-based NoCs. Based on the symmetry of the mesh, tiles are grouped into partitions, and then nodes with high communication volumes are mapped first to the partitions. Addo-Quaye et al. [7] proposed a thermal-aware 3D mapping technique based on genetic algorithm.

Murali et al. [56] proposed Nmap, a three-phase algorithm to find near optimal mapping solution. However, these algorithms have a very high computational complexity. Wang et al. [57] proposed a mapping algorithm for 3D NoCs based on run-time incremental mapping technique [58]. Here, the algorithm tries to map applications to convex regions while utilizing as many vertical links as possible in the mapping process. This approach increases the number of 3D routers (vertical links) which are costly and have high power consumption.

Hence in [59] a reliability and power-aware technique for mapping multiple applications to 3D NoCs is presented. The proposed technique automatically generates low latency heterogeneous 3D NoCs by evaluating the number and placement of 2D and 3D routers required as well as the communication characteristics of tasks assigned to the processing elements. The mapping algorithm has been evaluated and compared with existing heuristic mapping algorithms (CastNet, Onyx and Nmap), optimal mapping (branch-and-bound) and a random mapping with various realistic traffic patterns. Experimental results show that NoCs mapped with the proposed algorithm in [59] have lower energy consumption and significant reduction in packet delays compared to other algorithms.

7. RESEARCH OPPORTUNITIES

Various proposal and designs have been studied in this paper. However, there are still various potential issues that need to be investigated. In this section, several current challenges as well as possible extensions are discussed as a motivation for future research.

Most of the previous and recent literature on minimizing the number of TSVs has focused on combining 2D and 3D routers in NoCs in a given topology, typically mesh. The effect of modifying topology in order to reduce the manufacturing cost while maintaining network performance needs more investigation. Considering the design of a hybrid interconnection paradigm that incorporates an appropriate mix of different NoC topologies, such as meshes (3D and 2D), buses, stacked torus, point-to-point in terms of area, latency and power efficiency, and efficient routing techniques need to be proposed. A possible research topic could include the exploration and implementation of a generic solution for 3D NoC technology and performance constraints that incorporates the best topology that resolves the issue of interlayer connectivity. For example, the investigation can present critical analysis of heterogeneous schemes and propose new combination of NoC topologies such as, semi-torus 3D NoC, Mesh-Stacked torus, Diagonal Bus-3D mesh NoC and 3D NoC-bus hybrid (such as Bus backbone, H-tree Mesh combinations, point-to-point and torus combinations), etc. In addition, this research topic can be extended to investigate various mapping algorithms to effectively exploit the architecture variation of new hybrid 3D NoCs.

Another outstanding challenge of 3D NoC in application-specific design is issue of assigning processors and switches to specific locations on the IC layers without incorporating complex algorithms. A significantly improved throughput is expected if processors and switches are efficiently assigned to the IC layers, however, issues such as heat dissipation, floorplanning and TSV placement needs to be addressed. Extending this research gap to heterogeneous 3D NoC architectures, efficient floorplan-aware mapping algorithms and performance trade-off associated with exploring new partitions is an open area for investigation. For example, cores with high bandwidth demands can be put in one partition and connected to high performance routers whiles, cores with lower bandwidth requirements can be grouped in another partition that is has a 3D NoC-bus hybrid router. Various issues such as how best the interlayer connections can be achieved and the effect of the on-chip communication infrastructure on the core assignment and the floor-planning can be examined.

Heterogeneous 3D NoC architectures have limited number of vertical links. This implies that a faulty vertical link could have a significant impact on the performance of the network if efficient error reporting and fault tolerant routing algorithms or reconfigurable architectures are not employed. Most of the existing fault tolerant schemes employ TSV redundancy or virtual channels. On the other hand, this contradicts with the main aim of reducing the number of vertical links by introducing heterogeneity. Also, the introduction of fault-aware algorithms will have an impact on the area, power consumption and the clock frequently of the router architecture. Also, the issue of how to efficiently reroute packets without sacrificing the performance or creating deadlocks in the heterogeneous 3D NoC architecture is open for research. A possible solution will be to implement a limited number of fault tolerant nodes based on the evaluation of fault probability of the network. However, this needs to be heavily investigated as the location of the fault and the response time of the fault will impact the NoCs performance. Also, issues such the appropriate number of TSVs in terms of cost for a given 3D NoC can be estimated to evaluate the right number and placement of spare TSVs for the purpose of error detection and correction.

Another drawback to 3D NoCs technology is the unavailability of CAD tools. The insufficiency of physical design tools that incorporates TSVs and 3D die stacking inhibits the acceptance of this technology into the mainstream. Recently, a commercial tool for TSV-based 3D IC design, MAX-3D Layout Editor was made available by Micro Magic, Inc [60]. However, the support feature of this tool is limited to only layout editing for 3D ICs. More importantly, it does not offer automatic placement and routing. In addition, even in the commercial world, there is no tool available for timing, signal integrity, power consumption, power supply noise, and design cost analysis of TSV and 3D stacking without the need of a third party tool [19]. Design impact and implementation of such tools is a significant research topic to be investigated.

8. CONCLUSION

This paper presents a study into the issue of manufacturing cost of 3D NoCs with the main objective of reducing the number of TSVs, area and power consumption while maintaining the performance of the network. This objective has been handled through various techniques of mitigating the manufacturing cost of high performance NoCs in various literature which have been discussed in the paper. In summary, existing contributions on 3D NoCs in area of heat dissipation, floorplanning, switch placement, NoC partitioning, router area and power consumption have been presented and the research gaps are identified to as a motivation. Consequently, the concept of 3D NoCs is introduced along with the basic principles to emphasize on the research gaps in several yield, area, manufacturing, performance and simulation of 3D NoC technology. As a result, a study of various heterogeneous 3D NoC architectures by combining 2D and 3D routers, associated routing algorithms and design automation techniques have been provided. Future work includes an exploration and implementation of a generic solution for 3D NoC technology and performance constraints that incorporates the best topology that resolves the issue of interlayer connectivity.

9. REFERENCES

[1] P. P. Pande, C. Grecu, M. Jones, A. Ivanov, and R. Saleh, "Performance evaluation and design trade-offs for network-on-chip interconnect architectures," *IEEE Transactions on Computers*, vol. 54, no. 8, pp. 1025–1040, 2005.

[2] E. Salminen, A. Kulmala, and T. D. Hmlinen, "Survey of Network-on-chip Proposals," *White paper OCP-IP*, vol. 1, 2008.

[3] I. Loi, F. Angiolini, and L. Benini, "Supporting vertical links for 3D networks-on-chip: toward an automated design and analysis flow," in *Proc. of the 2nd international conference on Nano-Networks*, 2007, pp. 1 – 5.

[4] G. D. Micheli and L. Benini, *Networks on Chips: Technology and Tools*. Morgan Kaufmann, First Edition, 2006.

[5] V. Pavlidis and E. Friedman, "3-D Topologies for Networks-on-Chip," *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, vol. 15, no. 10, pp. 1081–1090, 2007.

[6] B. Feero and P. P. Pande, "Performance Evaluation for Three-Dimensional Networks-On-Chip," in *Proceedings of the IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, 2007, pp. 305–310.

[7] C. Addo-Quaye, "Thermal-aware mapping and placement for 3-D NoC designs," 2005, pp. 25–28.

[8] J. Kim, C. Nicopoulos, D. Park, R. Das, Y. Xie, V. Narayanan, M. S. Yousif, and C. R. Das, "A novel dimensionally-decomposed router for on-chip communication in 3D architectures," *SIGARCH Comput. Archit. News*, vol. 35, no. 2, pp. 138–149, 2007.

[9] S. Lim, "Physical design for 3D system on package," *Design Test of Computers, IEEE*, pp. 532–539, 2005.

[10] C. Seiculescu, S. Murali, L. Benini, and G. De Micheli, "SunFloor 3D: A tool for Networks On Chip topology synthesis for 3D systems on chips," 2009, pp. 9–14.

[11] K. Siozios, A. Bartzas, and D. Soudris, "Three dimensional network-on-chip architectures," in *Networks-on-Chips: Theory and Practice*, H. E. Fayez Gebali and M. W. El-Kharashi, Eds. CRC Press, 2009, pp. 1–28.

[12] S. Stergiou, F. Angiolini, S. Carta, L. Raffo, D. Bertozzi, and G. D. Micheli, "Xpipes Lite: A synthesis oriented design library for networks on chips," in *Design, Automation and Test in Europe (DATE)*. IEEE, 2005, pp. 1188–1193.

[13] 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.

[14] S. K. Lee, C. Y. Kang, P. Kirsch, S. Arkalqud, R. Jammy, and B. H. Lee, "Comprehensive study for rf interference limited 3d tsv optimization," in *International Symposium on VLSI Technology, Systems, and Applications (VLSI-TSA)*, 2013, pp. 1–2.

[15] P. C. Chew, L. Li, J. Xue, and W. Eklow, "Through silicon via (tsv) redundancy - a high reliability, networking product perspective," in *International Conference on Electronic Materials and Packaging (EMAP)*, 2012, pp. 1–5.

[16] V. F. Pavlidis and E. G. Friedman, *Three-dimensional Integrated Circuit Design*. Morgan Kaufmann Publishers Inc., 2009.

[17] B. Feero and P. Pande, "Networks-on-Chip in a Three-Dimensional Environment: A Performance Evaluation," *IEEE Transactions on Computers*, vol. 58, no. 1, pp. 32 –45, 2009.

[18] R. Patti, "Three-dimensional integrated circuits and the future of system-on-chip designs," *Proceedings of the IEEE*, vol. 94, no. 6, pp. 1214 –1224, 2006.

[19] S. K. Lim, "TSV-Aware 3D Physical Design Tool Needs for Faster Mainstream Acceptance of 3D ICs," *ACM DAC Knowledge Center (dac.com)*, 2010.

[20] D. Velenis, M. Stucchi, E. Marinissen, B. Swinnen, and E. Beyne, "Impact of 3d design choices on manufacturing cost," in *IEEE International Conference on 3D System Integration (3DIC)*, 2009, pp. 1 – 5.

[21] C. Feng, M. Zhang, J. Li, J. Jiang, Z. Lu, and A. Jantsch, "A low-overhead fault-aware deflection routing algorithm for 3d network-on-chip," in *IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, 2011, pp. 19–24.

[22] C. Liu, L. Zhang, Y. Han, and X. Li, "Vertical interconnects squeezing in symmetric 3D mesh Network-on-Chip," in *Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2011, pp. 357 –362.

[23] "Tour guide to 3d-ic design tools and services," http://www.gsaglobal.org/eda/docs/, 2012, online; accessed June-2012.

[24] I. T. R. for Semiconductors (ITRS), http://www.itrs.net/, 2013, [Online; accessed February].

[25] G. H. Loh, Y. Xie, and B. Black, "Processor design in 3d die-stacking technologies," *IEEE-Micro*, vol. 27, pp. 31 –48, 2007.

[26] D. H. Kim, S. Mukhopadhyay, and S. K. Lim, "Through-silicon-via aware interconnect prediction and optimization for 3d stacked ics," in *International workshop on System level interconnect prediction*, 2009, pp. 85–92.

[27] D. H. Kim, S. Mukhopadhyay, and S.-K. Lim, "Tsv-aware interconnect length and power prediction for 3d stacked ics," in *IEEE International Interconnect Technology Conference (IITC)*, 2009, pp. 26–28.

[28] D. H. Kim, K. Athikulwongse, and S.-K. Lim, "A study of through-silicon-via impact on the 3d stacked ic layout," in *IEEE/ACM International Conference on Computer-Aided Design - Digest of Technical Papers (ICCAD).*, 2009, pp. 674–680.

[29] R. Weerasekera, L.-R. Zheng, D. Pamunuwa, and H. Tenhunen, "Extending systems-on-chip to the third dimension: performance, cost and technological tradeoffs," in *IEEE/ACM international conference on Computer-aided design (ICCAD)*, 2007, pp. 212–219.

[30] A. K. Mishra, N. Vijaykrishnan, and C. R. Das, "A case for heterogeneous on-chip interconnects for CMPs," in *Annual International Symposium on Computer Architecture (ISCA)*, 2011, pp. 389–400.

[31] M. Al Faruque and J. Henkel, "Minimizing virtual channel buffer for routers in on-chip communication architectures," in *Design, Automation and Test in Europe (DATE)*, march 2008, pp. 1238 –1243.

[32] A. Sharifi and H. Sarbazi-Azad, "Power consumption and performance analysis of 3d nocs," in *Asia-Pacific Computer Systems Architecture Conference*, 2007, pp. 209–219.

[33] T. Xu, P. Liljeberg, and H. Tenhunen, "A study of Through Silicon Via impact to 3D Network-on-Chip design," in *2010 International Conference On Electronics and Information Engineering (ICEIE)*, 2010, pp. 333 – 337.

[34] F. Li, C. Nicopoulos, T. Richardson, Y. Xie, V. Narayanan, and M. Kandemir, "Design and Management of 3D Chip Multiprocessors Using Network-in-Memory," in *International Symposium on Computer Architecture (ISCA)*, 2006, pp. 130–141.

[35] M. O. Agyeman, A. Ahmadinia, and A. Shahrabi, "Heterogeneous 3d network-on-chip architectures: area and power aware design techniques," *Journal of Circuits, Systems and Computers*, vol. 22, no. 4, p. 1350016, 2013.

[36] ——, "Low power heterogeneous 3d networks-on-chip architectures," in *International Conference on High Performance Computing and Simulation (HPCS)*, 2011, pp. 533 –538.

[37] M. O. Agyeman and A. Ahmadinia, "Optimising heterogeneous 3d networks-on-chip," in *PARELEC*, 2011, pp. 25–30.

[38] H. Ying, A. Jaiswal, T. Hollstein, and K. Hofmann, "Deadlock-free generic routing algorithms for 3-dimensional networks-on-chip with reduced vertical link density topologies," *Journal of Systems Architecture*, vol. 59, no. 7, pp. 528 – 542, 2013.

[39] A.-M. Rahmani, K. Latif, P. Liljeberg, J. Plosila, and H. Tenhunen, "A stacked mesh 3d noc architecture enabling congestion-aware and reliable inter-layer communication," in *Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*, 2011, pp. 423 –430.

[40] M. O. Agyeman, A. Ahmadinia, and A. Shahrabi, "Efficient routing techniques in heterogeneous 3d networks-on-chip," *Parallel Computing*, vol. 39, no. 9, pp. 389–407, 2013.

[41] M. O. Agyeman and A. Ahmadinia, "An adaptive router architecture for heterogeneous 3d networks-on-chip," in *International IEEE NORCHIP Conference*, 2011, pp. 1 –4.

[42] D. Park, S. Eachempati, R. Das, A. K. Mishra, Y. Xie, N. Vijaykrishnan, and C. R. Das, "Mira: A multi-layered on-chip interconnect router architecture," in *Annual International Symposium on Computer Architecture*, 2008, pp. 251–261.

[43] S. Pasricha, "Exploring serial vertical interconnects for 3d ics," in *ACM/IEEE Design Automation Conference (DAC)*, 2009, pp. 581 –586.

[44] ——, "A framework for tsv serialization-aware synthesis of application specific 3d networks-on-chip," in *International Conference on VLSI Design (VLSID)*, 2012, pp. 268 –273.

[45] W. Liwei, C. Yang, L. Xiaohui, and Z. Xiaohu, "Application specific buffer allocation for wormhole routing networks-on-chip," pp. 37–42, 2008.

[46] J. Hu, U. Y. Ogras, and R. Marculescu, "System-level buffer allocation for application-specific networks-on-chip router design," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 25, no. 12, pp. 2919 –2933, 2006.

[47] S. Foroutan, Y. Thonnart, R. Hersemeule, and A. Jerraya, "An analytical method for evaluating network-on-chip performance," in *Design, Automation Test in Europe Conference Exhibition (DATE)*, march 2010, pp. 1629 –1632.

[48] A. S. Kumar, M. P. Kumar, S. Murali, V. Kamakoti, L. Benini, and G. D. Micheli, "A buffer-sizing algorithm for network-on-chips with multiple voltage-frequency islands," *J. Electrical and Computer Engineering*, vol. 2012, 2012.

[49] T. C. Xu, G. Schley, P. Liljeberg, M. Radetzki, J. Plosila, and H. Tenhunen, "Optimal placement of vertical connections in 3d network-on-chip," *Journal of Systems Architecture*, vol. 59, no. 7, pp. 441 – 454, 2013.

[50] M. O. Agyeman and A. Ahmadinia, "A systematic generation of optimized heterogeneous 3d networks-on-chip architecture," in *NASA/ESA Conference on Adaptive Hardware and Systems (AHS)*, 2013, pp. 79–83.

[51] C. Ababei, H. S. Kia, O. P. Yadav, and J. Hu, "Energy and reliability oriented mapping for regular networks-on-chip," in *NOCS*, 2011, pp. 121–128.

[52] 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.

[53] X. Wang, M. Yang, Y. Jiang, and P. Liu, "A power-aware mapping approach to map ip cores onto nocs under bandwidth and latency constraints," *TACO*, vol. 7, no. 1, 2010.

[54] M. Janidarmian, A. Khademzadeh, and M. Tavanpour, "Onyx: A new heuristic bandwidth-constrained mapping of cores onto tile-based Network on Chip," *Ieice Electronic Express*, vol. 6, pp. 1–7, 2009.

[55] S. Tosun, "New heuristic algorithms for energy aware application mapping and routing on mesh-based nocs," *Journal of Systems Architecture*, vol. 57, no. 1, pp. 69 – 78, 2011.

[56] S. Murali and G. De Micheli, "Bandwidth-constrained mapping of cores onto NoC architectures," in *Design, Automation and Test in Europe Conference and Exhibition (DATE)*, 2004, pp. 896 – 901.

[57] X. Wang, M. Palesi, M. Yang, Y. Jiang, M. C. Huang, and P. Liu, "Power-aware run-time incremental mapping for 3-d networks-on-chip," in *NPC*, 2011, pp. 232–247.

[58] C.-L. Chou and R. Marculescu, "Run-time task allocation considering user behavior in embedded multiprocessor networks-on-chip," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 29, no. 1, pp. 78 –91, 2010.

[59] M. Agyeman and A. Ahmadinia, "Optimised application specific architecture generation and mapping approach for heterogeneous 3d networks-on-chip," in *IEEE International Conference on Computational Science and Engineering (CSE)*, 2013, pp. 794–801.

[60] M. M. Inc, http://www.micromagic.com, 2010, [Online; accessed July].