Low-Latency Wireless 3D NoCs via Randomized Shortcut Chips

Hiroki Matsutani[1,2], Michihiro Koibuchi[2], Ikki Fujiwara[2], Takahiro Kagami[1], Yasuhiro Take[1],
Tadahiro Kuroda[1], Paul Bogdan[3], Radu Marculescu[4], and Hideharu Amano[1,2]

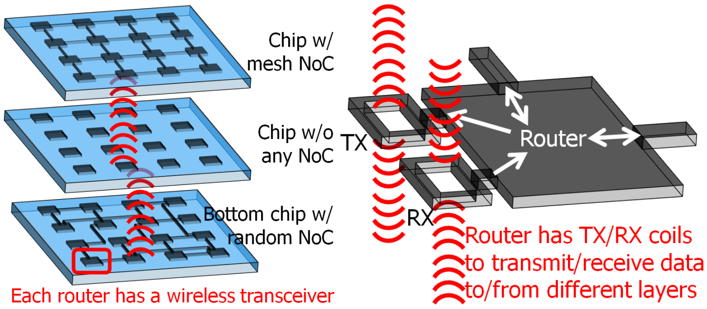
[1]Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Japan, cube@am.ics.keio.ac.jp
[2]National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan, {koibuchi,ikki}@nii.ac.jp
[3]University of Southern California, 3740 McClintock Ave., Los Angeles, USA, pbogdan@usc.edu
[4]Carnegie Mellon University, 5000 Forbes Ave., Pittsburgh, USA, radum@cmu.edu

***Abstract*—In this paper, we demonstrate that we can reduce the communication latency significantly by inserting a fraction of randomness into a wireless 3D NoC (where CMOS wireless links are used for vertical inter-chip communication) when considering the physical constraints of the 3D design space. Towards this end, we consider two cases, namely 1) replacing existing horizontal 2D links in a wireless 3D NoC with randomized shortcut NoC links and 2) enabling full connectivity by adding a randomized NoC layer to a wireless 3D platform with partial or no horizontal connectivity. Consequently, the packet routing is optimized by exploiting both the existing and the newly added random NoC. At the same time, by adding randomly wired shortcut NoCs to a wireless 3D platform, a good balance can be established between the modularity of the design and the minimum randomness needed to achieve low latency, and experimental results show that by adding a random NoC chip to wireless 3D CMPs without built-in horizontal connectivity, the communication latency can be reduced by as much as 26.2% when compared to adding a 2D mesh NoC. Also, the application execution time and average flit transfer energy can be improved accordingly.**

I. Introduction

Recent advances in semiconductor technology allow us to integrate many processing cores on a single chip. These many-core processors can be used for embedded high-performance applications and cloud computing. 3D integration is a promising approach to integrate more cores. Indeed, multiple dies (or wafers) that integrate processors and memory modules can be stacked vertically in a single package to reduce the wire length, increase the memory bandwidth, and improve performance. For such chips, the 3D NoC architecture has been used to connect cores on different dies (using inter-chip network) in addition to those residing on the same die (using intra-chip networks). Various 3D integration technologies are available: micro-bump, wireless (e.g., capacitive- and inductive-coupling) between stacked dies, and through-silicon via (TSV) between stacked wafers [1]. Although TSV-based 3D IC design is becoming more mature, in this paper we focus on the more challenging wireless inductive-coupling option [2][3], as it offers both scalability to stack more than two chips and flexibility to connect known-good-dies selected after chip fabrication.

Figure 1 illustrates an example of a wireless 3D NoC that stacks three chips in a System-in-Package. As shown, every router has CMOS TX/RX coils for vertical communication (e.g., bus or point-to-point), in addition to the horizontally wired links needed to connect the cores or to other routers on the same die.



(a) Wireless 3D NoC (b) Router with TX/RX coils

Fig. 1. Wireless 3D system consisting of three chips. Although all these chips have vertical links at the pre-specified locations, they may or may not have horizontal NoCs. In this example, top and bottom chips have mesh and random NoCs respectively while the middle one does not.

In this paper, we demonstrate that, by adding random connectivity via wireless 3D NoCs, we can significantly reduce communication latency while incurring minimum design complexity. To this end, we examine two cases: 1) replacing the existing horizontal 2D NoCs in a wireless 3D NoC with random shortcut NoCs and 2) adding a random NoC chip, in which the horizontally wired links are randomly determined, to a wireless 3D platform with partial or no horizontal NoCs in order to achieve full connectivity (Figure 1(a) shows the latter case). The random NoC chip has redundant horizontal links (in addition to the regular links) connected via routers and FPGA-like switch boxes on the same die. By reconfiguring the switch boxes randomly, we can build a random NoC chip with a unique horizontal wiring pattern.

In such a platform, the packets routing can be optimized by exploiting both the newly added random NoCs and the existing regular NoCs by using the irregular up*/down* routing [4] relying on the spanning trees optimization method proposed for wireless 3D NoCs [5].

We note that using randomness to minimize the hop count and communication latency has been studied for interconnection networks of high-performance computing [6], datacenters [7], and planar 2D NoCs [8][9][10]. However, the 3D NoC topologies are more challenging because they are physically restricted due to vertical link locations; for example, a vertical link is formed only between two end points (e.g., TSVs, micro-bumps, and inductors) implemented on exactly same horizontal location on different dies (or wafers), as shown in Figure 1(a).

978-3-9815370-2-4/DATE14/©2014 EDAA

From an engineering point of view, a wireless 3D IC system is a collection of dies, each of which is developed more or less independently. For example, dies built upon different process technologies, such as processor die and DRAM die, can be designed independently and stacked only later. Thus, adding randomly wired shortcut NoCs to wireless 3D systems ensures a good balance between modular design and minimum design complexity for low latency.

The remaining of this paper is organized as follows. Section II surveys low latency network topologies and wireless 3D NoCs. Section III provides a detailed latency analysis of the randomly wired shortcut chips. Section IV validates the analysis results based on a case study of wireless 3D CMPs. Finally, Section V summarizes our main findings.

II. Background and Related Work

A. On-Chip Small-World Networks

Traditionally, the usefulness of random shortcuts has been noted for complex networks, e.g., social networks and the Internet [11]. The scale-free and clustering properties in small-world networks lead to low diameter and average shortest path length, and also provide robustness to random edge removal. Consequently, researchers have designed approaches that exploit the small-world property in the areas of high-performance computing networks [6], datacenter networks [7], peer-to-peer overlay networks, or wireless sensor networks.

Small-world effects have been exploited for on-chip networks and high-performance computing networks to reduce their hop counts. By introducing long-range links, a small-world topology that adds wired shortcut links to k-ary 2-mesh can be built in order to reduce the hop counts [8]. In addition to conventional wired on-chip networks, small-world wireless 2D NoCs that employ millimeter-wave have been studied recently [9]. A run-time macro-scale topology reconfiguration, called Skip-links, has been also proposed to dynamically exploit the small-world effects of on-chip networks [10].

Starting from these ideas, this paper focuses on the small-world effects on wireless 3D NoCs, in which vertical randomness is physically restricted. We examine two cases: 1) replacing the existing horizontal 2D NoCs in a wireless 3D NoC with random shortcut NoCs and 2) adding a random NoC chip to wireless 3D systems that have partial or no horizontal NoC links to make full connectivity.

As seen above, using the enormous amount of wire resources enables us to consider NoCs that have links longer than the length of each tile. Thus, a random topology, in which random shortcut links are picked so that they do not lead to a very long length, is a practical choice for low latency.

B. Vertical Communication Schemes

Inductive-coupling [1][2][3] is a die-level wireless interconnection scheme that uses square coils as data transmitters. The coils can be implemented with common metal layers of the chip; thus no special process technology is required. By stacking chips in a package, the communication distance between them can be reduced to several tens of micrometers (e.g., 40um for chip thickness and 10um for glue). By increasing the number of coil turns, the transmission gain can be increased and the communication distance becomes longer than the chip thickness. Thus, more than two known-good-dies can be connected by using a face-to-back connection.

The inductive-coupling can enable the customization of hardware components (or chips) in a package to satisfy the application requirements at a low cost. That is, we can add, remove, and swap the known-good-dies (e.g., processor and memory chips) in a package to meet the requirements without making new mask patterns. As the wireless inductive-coupling enables us to stack additional chips afterward, in this paper we demonstrate that adding random connectivity via wireless 3D NoCs can reduce communication latency with minimum design complexity.

A wireless vertical link can be implemented with a shared bus or multiple point-to-point (P2P) vertical links, as reported in [12]. Although the bus structure intrinsically does not offer much scalability when the number of communication points (e.g., number of chips) increases, it is an efficient way of connecting a moderate number of dies or wafers in 3D CMPs. Dynamic Time Division Multiple Access (dTDMA) [13][14] bus is introduced for the distribution of bus mastership between multiple chips, while point-to-point NoCs are used inside each chip. Both vertical buses and P2P links are implemented with the inductive-coupling and either one can be used depending on the number of chips stacked and communication pattern, such as unicast- or multicast-based communication.

III. Design Space Exploration of Wireless 3D Architectures

In this section, we examine and analyze various design parameters for adding randomness to a wireless 3D NoC in terms of zero-load latency. We create a random topology with random links picked so that they will not be longer than the specified maximum long-range link length [1]. Then, the distance between two tiles is computed using the Manhattan distance. The default setting of parameters is shown in Table I.

TABLE I. Initial parameters in latency analysis.

# of tiles per chip × # of chips	16 × 4
Horizontal degree	4
Total # of chips, # of random chips	4, 2
Maximum random link length	2 tiles
On-chip non-random topology	k-ary 2-mesh

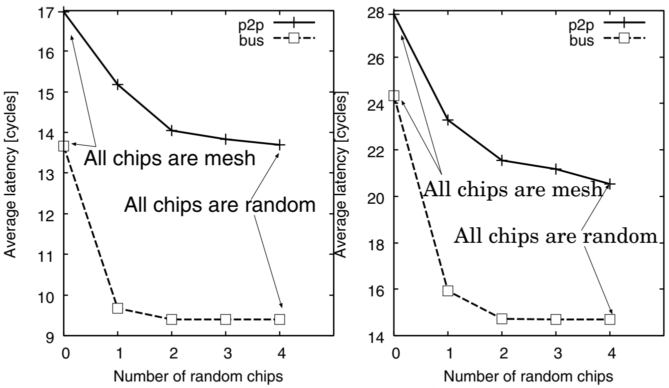
Assuming that a packet that consists of L flits (including a single header flit) goes through h wormhole routers and link bandwith is BW, its zero-load latency is calculated as:

$$T = T_{lt}(h-1) + T_{rt}h + L/\mathrm{BW}, \qquad (1)$$

where T_{rt} and T_{lt} are the latencies for packet forwarding at a router and link traversal, respectively.

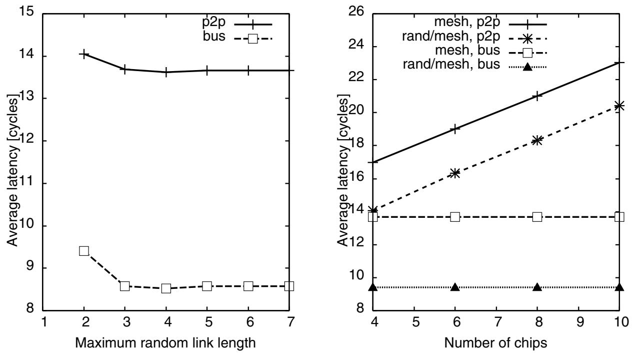
The router latency is set to three cycles. The horizontal link latency is selected based on the distance between two end points. Using repeated on-chip wires, a single cycle is assumed for 1- and 2-hop distances (e.g., k-ary 2-meshes and folded k-ary 2-tori), while two cycles are assumed for longer distances. In this analysis, a random long-range link is set

[1] Here, we define the horizontal degree of a router as the number of links needed to connect it to the other routers on the same chip.



(a) 64 tiles (four 16-tile chips) (b) 256 tiles (four 64-tile chips)

Fig. 2. Zero-load latency vs. number of random chips for 3D NoCs consisting of random and k-ary 2-mesh chips.



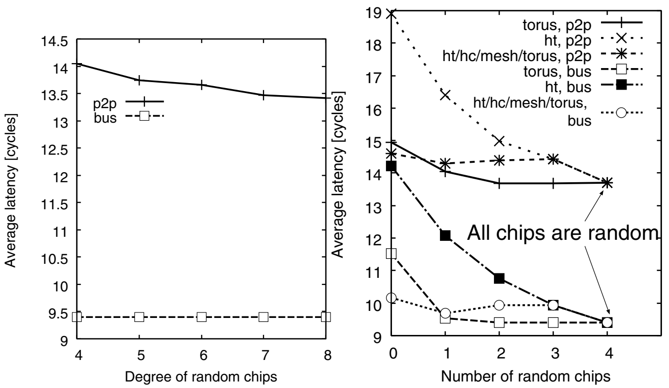
(a) Maximum random link length (b) Number of chips

Fig. 3. Zero-load latency vs. (a) maximum random link length and (b) number of chips for 3D NoCs consisting of random and k-ary 2-mesh chips.

to bypass up to two tiles by default. As for the wireless vertical link, a single cycle is assumed for communicating between neighboring chips using a P2P link. A single cycle is also assumed for vertical communication between any pair of source and destination chips when a dTDMA bus is used.

a) Number of Random Chips: Here, we explore the number of random chips needed in 3D NoCs that consist of four chips. Figure 2 shows the relationship between the zero-load latency and the number of our random topology chips in which the maximum random-link length cannot exceed twice the length of a normal short link. We attempt both cases for the P2P links and the dTDMA buses for the vertical connections. When the x-axis reaches four, all the chips take the random topology. One or two random chips significantly reduce the zero-load latency, while adding more random chips gracefully decreases it. Our recommendation is thus the use of one or two random chips in the 3D NoCs.

b) Maximum Length of Random Link: We are also interested in the maximum length of random links. Figure 3(a) shows the average zero-load latency vs. maximum random link length for 3D NoCs with P2P links and dTDMA buses, in which two chips have random topology, respectively. We set the maximum random link length as a parameter for the random topology generation. That is, two nodes are randomly picked up, and, if their distance is less than or equal to the maximum random link length, they are connected with a link. To fully exploit the random effect to shorten the zero-load latency, Figure 3(a) illustrates that the longest random link should be of length two or three; this is because the latency reduction when using long-range links longer than three tiles cannot be observed. Considering these observations, bypassing two tiles is an economical choice when exploiting randomness; consequently, it is the default configuration for our random topology chip.



(a) Horizontal degree of random chips (b) Baseline chip topologies

Fig. 4. Zero-load latency vs. (a) horizontal degree of random chips and (b) different baseline chip topologies for 3D NoCs.

c) Number of Chips: In what follows, we discuss the performance impact when changing the total number of chips in 3D NoCs. Figure 3(b) plots the zero-load latency vs. total number of chips for four 3D NoC configurations. "mesh, p2p" is a 3D NoC, in which each chip has k-ary 2-mesh topology connected with P2P vertical links. It thus forms k-ary 3-mesh. We consider both dTDMA buses and P2P links for vertical connections; thus, "mesh, bus" employs dTDMA buses for vertical connections. On the other hand, "rand/mesh, p2p" consists of two random chips and the remainder are k-ary 2-mesh chips connected with P2P vertical links. "rand/mesh, bus" employs dTDMA buses for vertical connections. We found that our random chips improve the zero-load latency by 31% and 17% with dTDMA buses and P2P links, respectively. In particular, in the case of dTDMA buses, a larger number of paths go through our random topology chips; thus this reduction becomes larger when compared to P2P links. As the number of chips increases, the effect of our random topology becomes relatively small in the case of P2P links, whereas its good improvement can be observed constantly in the case of dTDMA buses.

d) Horizontal Degree of Random Topology: Here, we analyze the zero-load latency when the horizontal degree of random chips is increased, while that of the remaining non-random chips is fixed to four. In this case, as the horizontal degree of random chips increases, the zero-load latency decreases very slightly in the case of P2P links, as shown in Figure 4(a). When considering the balance between the cost and latency reduction, we mainly focus on the case for four horizontal degrees in our random topology.

e) Different Baseline Non-Random Topology: We have evaluated our random topology chip for 3D NoCs in which the remaining chips employ k-ary n-mesh. Hereby, we evaluate 3D NoC topology variations that include different non-random on-chip 2D topologies. Figure 4(b) plots the zero-load latencies for 3D NoCs in which the baseline (non-random topology) is folded torus, H-tree, and hypercube (hc). The "ht/hc/torus/mesh" represents a 3D NoC that consists of hypercube, k-ary 2-mesh, folded k-ary 2-torus, and H-tree. Since the hypercube has a high degree ($\log_2 N$), its average link length becomes long as the number of tiles per chip increases. By contrast, our random topology chips maintain the average link length, yet still achieve the reduced zero-load latency. Obviously, by employing our random chips, the zero-load latency drastically decreases in all the baseline non-random topologies. Thus, our random topology is beneficial for reducing link length.

f) Summary: Based on the findings in this section, we recommend the use of one or two random topology chips for reducing the zero-load latency but moderate link lengths for 3D NoCs. The maximum random link length is limited to two tiles and the horizontal degree of four is enough. The random topology chip is efficient for various combinations of non-random topology chips for 3D NoCs.

IV. Experimental Results

We assume 3D NoCs that consist of four 16-tile chips resulting in 64 tiles in total. As horizontal topologies, we consider 4-ary 2-mesh and random topologies whose horizontal degree is set to four. The maximum long-range link length is limited to two tiles. The other assumptions are consistent with those in the previous section.

We consider two cases, namely 1) replacing existing horizontal 2D NoCs in a wireless 3D NoC with random shortcut NoCs and 2) adding a random NoC chip to wireless 3D systems that have partial or no horizontal NoC links in order to make full connectivity. The former case is referred to as “P2P-based wireless 3D NoC,” while the latter one is “Bus-based wireless 3D NoC.” More specifically, the following six configurations are compared in terms of the communication latency, application execution time, and flit transmission energy.

- *mmmm* : P2P-based wireless 3D NoC that consists of four mesh chips.
- *mrrm* : P2P-based wireless 3D NoC that consists of two mesh chips and two random chips.
- *rrrr* : P2P-based wireless 3D NoC that consists of four random chips.
- —*m* : Bus-based wireless 3D NoC that consists of a single mesh chip; the remaining three chips do not have NoCs.
- —*r* : Bus-based wireless 3D NoC that consists of a single random chip; the remaining three chips do not have NoCs.
- *m*–*r* : Bus-based wireless 3D NoC that consists of a single mesh chip and a single random chip; the remaining two chips do not have NoCs.

A. Target CMP Architecture and Evaluation Environments

We assume shared-memory CMPs, in which each processor has private L1 data and instruction caches, while the unified L2 cache banks are shared by all the processors. Their placement affects application performance; we assume the wireless 3D CMPs illustrated in Figure 5 that consist of four chips, each of which has two processors (CPUs) and eight L2 cache banks. Four memory controllers are attached to the four corners of the bottom chip. These processors, L2 cache banks, and memory controllers are interconnected via on-chip routers. A directory-based cache coherence protocol that uses three message classes (or virtual channels) is running on the NoC.

Table II lists the processor and network parameters. We used a full-system CMP simulator gem5 [15] to simulate the wireless 3D CMPs. We modified a detailed network model of gem5 to accurately simulate the P2P-based and Bus-based wireless 3D NoCs with random topology chips. We use nine parallel programs from the OpenMP implementation of NAS Parallel Benchmarks (NPB) to evaluate the application performance in the cases of the six wireless 3D NoC configurations. The number of threads was set to eight since the number of processors in the target CMPs is eight.

TABLE II. Simulation parameters (processor, memory, NoC).

Processor architecture	x86-64
L1 I/D cache size, latency	32 KB (line:64B), 1 cycle
L2 cache bank size, latency	256 KB (assoc:4), 6 cycles
Memory size, latency	4 GB, 160 cycles
Router pipeline	3 cycles for router + 1 cycle for link
Buffer size, flit size	5 flits per VC, 128 bits
Cache coherency protocol	MOESI directory
# of VCs	3 (one VC for each message class)
Control / data packet size	1 flit / 5 flits

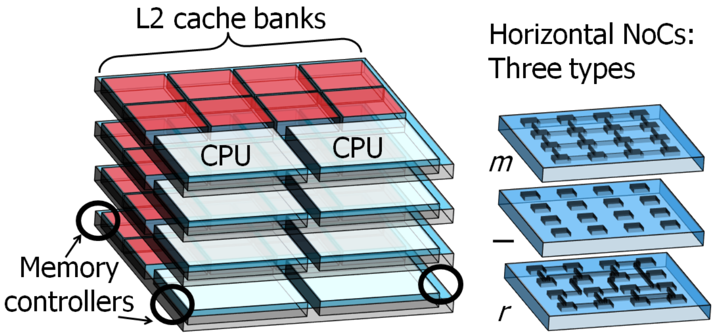


Fig. 5. Target 3D CMPs that consist of four chips, each of which has two CPUs, eight L2 cache banks, and a horizontal NoC. We examine three horizontal NoC types: m, $-$, and r (see right figure).

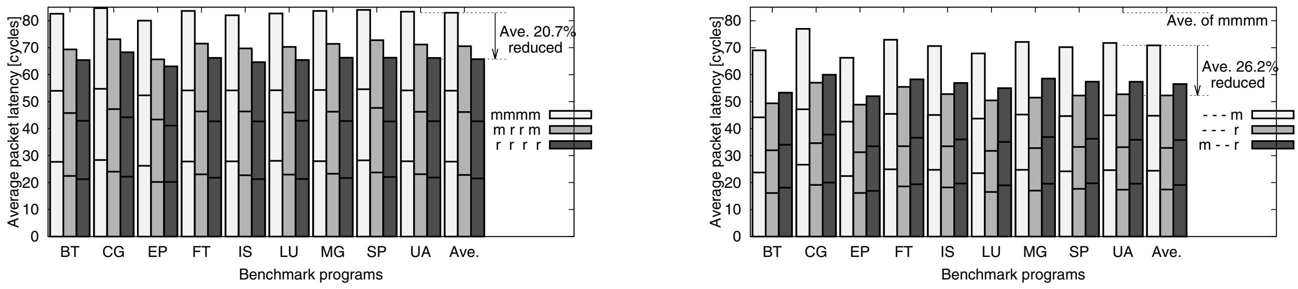
B. Topology and Routing Generation

When 3D NoCs have randomness, application performance varies depending on the generated random structures. To make comparisons fair and stable, we generate 1,000 topologies for each 3D NoC that includes random topology chips and calculate the average hop count for each of 1,000 topologies. Then, we pick up the most typical topology that has the closest hop count value to the average among the 1,000 topologies.

3D mesh NoCs, in which all the chips employ a mesh topology and are connected with P2P links vertically, use XYZ dimension-order routing to route packets. On the other hand, 3D NoCs that include random topology chips use an irregular routing algorithm, such as up*/down* routing [4]. In this experiment, we use the same routing strategy proposed in [5] to efficiently route packets in the 3D NoCs that have random topology chips. That is, the best spanning tree root that can minimize the hop count is selected for each message class or virtual channel layer. Then, routing paths are generated under the up*/down* rule with the selected spanning tree roots. Using this routing strategy, the packet routing can be optimized by exploiting both the existing regular NoCs and the newly added random NoCs, such as *m*–*r* case.

C. Performance Improvements

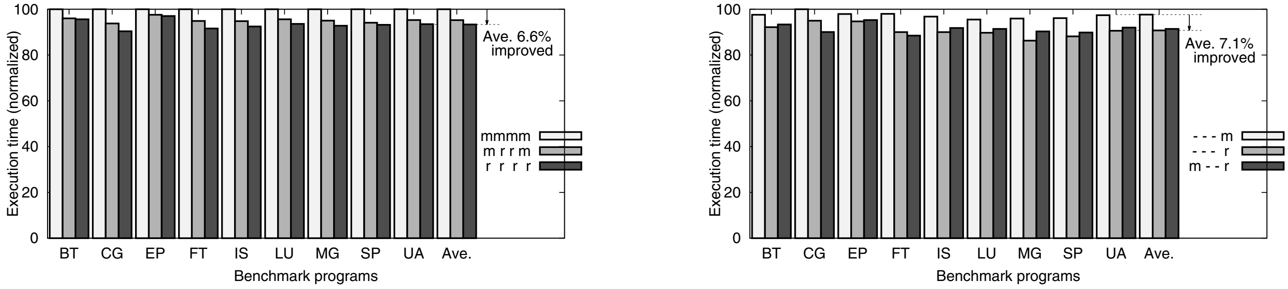
Figure 6(a) shows the communication latencies for nine NPB applications with *mmmm*, *mrrm*, and *rrrr* configurations.



(a) P2P-based wireless 3D NoCs

(b) Bus-based wireless 3D NoCs

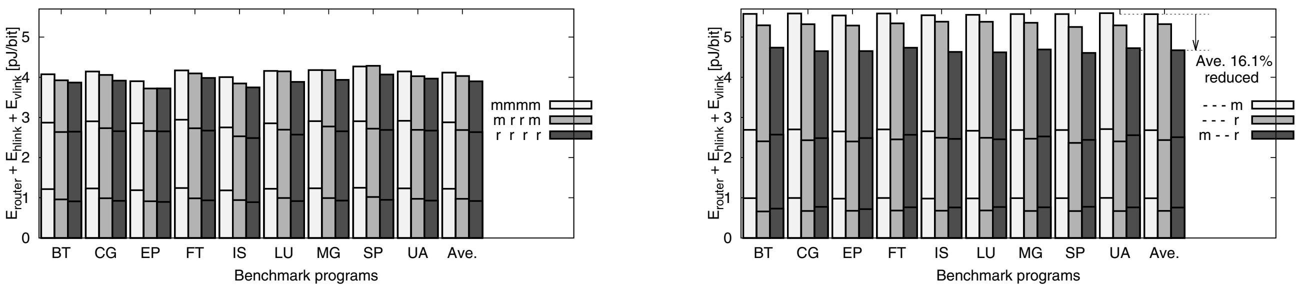
Fig. 6. Packet transfer latencies of wireless 3D NoCs. Each bar consists of packet latencies for message classes 0, 1, and 2, starting from the bottom.



(a) P2P-based wireless 3D NoCs

(b) Bus-based wireless 3D NoCs

Fig. 7. Application execution times of wireless 3D NoCs.



(a) P2P-based wireless 3D NoCs

(b) Bus-based wireless 3D NoCs

Fig. 8. Energy consumption of wireless 3D NoCs. Each bar consists of E_{router}, E_{hlink}, and E_{vlink}, starting from the bottom.

Each bar consists of packet latencies for message classes 0, 1, and 2. Their average is also shown in the right edge of the graph. As shown, the average communication latency of *mmmm* is by 20.7% shorter than that of *rrrr*. The average latency of *mrrm* is close to the *rrrr* configuration. Figure 6(b) shows the communication latencies of the Bus-based wireless 3D NoC configurations: *—m*, *—r*, and *m–r*. The average latency of *mmmm* is shown as a dashed line in the graph for comparison. The average latency of *—r* is 26.2% shorter than that of *—m*, and it is much less than that of *mmmm*. The average hop count of *m–r* is slightly longer than that of *—r*. This is because each node on the mesh chip (top chip) and the random chip (bottom chip) is connected to a single on-chip router while nodes on the non-NoC chips are directly connected to the vertical buses; thus, adding *m* to *—r* (i.e., *m–r*) increases hop counts of the paths which use the top chip.

Figure 7(a) shows the execution times of nine NPB applications with *mmmm*, *mrrm*, and *rrrr* configurations. The application execution times are normalized to that of *mmmm*. As shown, the application execution times of *mrrm* and *rrrr* are reduced compared to *mmmm*, as their communication latencies are reduced significantly as shown in Figure 6(a). Figure 7(b) shows the application execution times of *—m*, *—r*, and *m–r*. The application execution time results also reflect the communication latency reduction. This can prove beneficial for future complex heterogeneous computing platforms where deterministic and regularity of nodes may be lost.

D. Energy Consumption

To study the energy consumption per flit, we evaluate both the P2P-based and Bus-based wireless 3D NoCs that have random chips in terms of the average energy consumption when transmitting a single flit from the source to destination nodes. They are also compared against 3D meshes. This energy per flit can be estimated as:

$$\begin{aligned} E_{flit} &= w(E^{1hop}_{router} h_{router} + E^{1hop}_{hlink} h_{hlink} + E^{1hop}_{vlink} h_{vlink}) \\ &= w(E_{router} + E_{hlink} + E_{vlink}). \qquad (2) \end{aligned}$$

Here, w represents the flit-width. h_{router}, h_{hlink}, and h_{vlink} represent the number of router, horizontal link, and vertical link traversals on average. When we calculate h_{hlink}, a long-range link traversal that spans two tiles increments two, while a short link traversal between neighboring routers increments one, depending on the communication distance. E^{1hop}_{router}, E^{1hop}_{hlink}, and E^{1hop}_{vlink} correspond to the energy consumed by transmitting single bit data via a router, a horizontal 2mm link, and a vertical link (i.e., inductor). The E^{1hop}_{router} is set to 0.20pJ in this evaluation, based on the post-layout simulations of on-chip routers when a 65nm CMOS process with a 1.2V supply voltage was used[2]. The E_{hlink}^{1hop} is set to 0.43pJ, assuming that a semi-global interconnect whose wire capacitance was 0.20pF/mm (from ITRS 2007) was used for the 2mm horizontal links with repeaters inserted.

E_{vlink}^{1hop} is increased as the size of inductors increases depending on the communication distance. For a P2P-based wireless 3D NoC, the communication distance of an inductor is set to 50um (40um for chip thickness and 10um for glue between two chips). For a Bus-based wireless 3D NoC that stacks four chips, the communication distance of an inductor is three times longer than that of P2P-based one. Based on the circuit simulations, E_{vlink}^{1hop} of P2P-based wireless 3D NoC is set to 0.98pJ and that of Bus-based one is set to 1.925pJ.

We calculate the $(E_{router} + E_{hlink} + E_{vlink})$ values of the six configurations based on the hop count values (i.e., h_{router}, h_{hlink}, and ${}_{vlink}$) of all applications extracted from the full-system simulation results. Figure 8(a) shows the energy results of nine NPB applications with *mmmm*, *mrrm*, and *rrrr*, in which each bar represents its E_{router}, E_{hlink}, and E_{vlink} from the bottom. Figure 8(b) shows the flit energy results of Bus-based *—m*, *—r*, and *m–r* configurations. When we compare P2P-based and Bus-based wireless 3D approaches, the Bus-based approach consumes more energy since E_{vlink} of a vertical bus is twice larger than that of a P2P link (however, the Bus-based one is advantageous in terms of performance due to the shorter hop counts). Also, *m–r* topology reduces the energy consumption by 16.1% compared to *—m* topology, because *m–r* topology can reduce the utilization of vertical buses by using the horizontal mesh NoC of the top chip. In the case of *m–r*, the packet routing is optimized such that both the existing mesh NoC and the newly added random NoC are fully exploited to route packets.

V. Summary

In this paper, we have demonstrated that adding minimal random connectivity to wireless 3D NoCs can be beneficial in terms of hop count and communication latency. Toward this, we have examined two cases, namely 1) replacing existing horizontal 2D links in a wireless 3D NoC with randomized shortcut NoC links and 2) enabling full connectivity via adding a randomized NoC layer to a wireless 3D system with no or partial horizontal connectivity.

Our observations from a detailed latency analysis and experimental results using a full-system CMP simulators can be summarized as follows: First, using one or two random topology chips is beneficial for reducing the zero-load latency even though using moderate link lengths for 3D NoCs. For example, the maximum random link length can be limited to only two tiles. Only four ports are enough to make horizontal random links in each router. Second, adding a single random NoC chip to wireless 3D CMPs in which the remaining chips do not have any horizontal NoCs *reduces communication latency by 26.2%* compared to that of adding 2D mesh NoC. The application execution time and average flit transfer energy can also be improved accordingly.

Third, when the remaining chips have some horizontal NoCs, the packet routing can be optimized by exploiting both the newly added random NoC and existing regular ones.

Finally, we believe that adding random NoC chips to a wireless 3D system strikes a good balance between the modular design and low latency. However, more detailed scalability analysis is needed in terms of performance, cost, fault-tolerance, and routability, which is left for future work.

Acknowledgements A part of this work was supported by JSPS KAKENHI S Grant Number 25220002 and NII Research Fund. H.M. acknowledge supports from JST PRESTO and MIC SCOPE. R.M. was supported in part by the NSF grant CNS-1128624.

References

[1] W. R. Davis *et al.*, “Demystifying 3D ICs: The Pros and Cons of Going Vertical,” *IEEE Design and Test of Computers*, vol. 22, no. 6, pp. 498–510, Nov. 2005.

[2] N. Miura, H. Ishikuro, T. Sakurai, and T. Kuroda, “A 0.14pJ/b Inductive-Coupling Inter-Chip Data Transceiver with Digitally-Controlled Precise Pulse Shaping,” in *Proceedings of the International Solid-State Circuits Conference (ISSCC’07)*, Feb. 2007, pp. 358–359.

[3] S. Saito *et al.*, “MuCCRA-Cube: a 3D Dynamically Reconfigurable Processor with Inductive-Coupling Link,” in *Proceedings of the Field-Programmable Logic and Applications (FPL’09)*, Sep. 2009, pp. 6–11.

[4] M. D. Schroeder *et al.*, “Autonet: A High-speed, Self-configuring Local Area Network Using Point-to-point Links,” *IEEE Journal on Selected Areas in Communications*, vol. 9, pp. 1318–1335, Oct. 1991.

[5] H. Matsutani, P. Bogdan, R. Marculescu, Y. Take, D. Sasaki, H. Zhang, M. Koibuchi, T. Kuroda, and H. Amano, “A Case for Wireless 3D NoCs for CMPs,” in *Proceedings of the Asia and South Pacific Design Automation Conference (ASP-DAC’13)* , Jan. 2013, pp. 22–28.

[6] M. Koibuchi, H. Matsutani, H. Amano, D. F. Hsu, and H. Casanova, “A Case for Random Shortcut Topologies for HPC Interconnects,” in *Proceedings of the International Symposium on Computer Architecture (ISCA’12)*, 2012, pp. 177–188.

[7] A. Singla, C.-Y. Hong, L. Popa, and P. B. Godfrey, “Jellyfish: Networking Data Centers Randomly,” in *Proceedings of the USENIX conference on Networked Systems Design and Implementation (NSDI’12)*, Oct. 2012.

[8] U. Y. Ogras and R. Marculescu, “It’s a Small World After All: NoC Performance Optimization via Long-Range Link Insertion,” *IEEE Transactions on Very Large Scale Integration Systems*, vol. 14, pp. 693–706, 2006.

[9] A. Ganguly, K. Chang, S. Deb, P. P. Pande, B. Belzer, and C. Teuscher, “Scalable Hybrid Wireless Network-on-Chip Architectures for Multicore Systems,” *IEEE Transactions on Computers*, vol. 60, no. 10, pp. 1485–1502, Oct. 2011.

[10] S. J. Hollis, C. Jackson, P. Bogdan, and R. Marculescu, “Exploiting Emergence in On-chip Interconnects,” *IEEE Transactions on Computers*, Nov. 2012, (PrePrint).

[11] D. J. Watts and S. H. Strogatz, “Collective Dynamics of ‘Small-World’ Networks,” *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[12] H. Matsutani, Y. Take, D. Sasaki, M. Kimura, Y. Ono, Y. Nishiyama, M. Koibuchi, T. Kuroda, and H. Amano, “A Vertical Bubble Flow Network using Inductive-Coupling for 3-D CMPs,” in *Proceedings of the International Symposium on Networks-on-Chip (NOCS’11)* , May 2011, pp. 49–56.

[13] T. D. Richardson, C. Nicopoulos, D. Park, V. Narayanan, Y. Xie, C. Das, and V. Degalahal, “A Hybrid SoC Interconnect with Dynamic TDMA-Based Transaction-Less Buses and On-Chip Networks,” in *Proceedings of International Conference on VLSI Design (VLSID’06)*, Jan. 2006, pp. 657–664.

[14] M. O. Agyeman, A. Ahmadinia, and A. Shahrabi, “Low Power Heterogeneous 3D Networks-on-Chip Architectures,” in *Proceedings of the International Conference on High Performance Computing and Simulation (HPCS’11)*, Jul. 2011, pp. 533–538.

[15] N. Binkert *et al.*, “The gem5 Simulator,” *ACM SIGARCH Computer Architecture News*, vol. 39, no. 2, pp. 1–7, May 2011.

[2] These values depend on bit pattern of flits. We assume a random bit pattern.