3D NoCs with Active Interposer for Multi-Die Systems

Vasil Pano
Drexel University
3141 Chestnut St
Philadelphia, PA, USA 19104
vasilpano@gmail.com

Ragh Kuttappa
Drexel University
3141 Chestnut St
Philadelphia, PA, USA 19104
fr67@drexel.edu

Baris Taskin
Drexel University
3141 Chestnut St
Philadelphia, PA, USA 19104
taskin@coe.drexel.edu

ABSTRACT

Advances in interconnect technologies for system-in-package manufacturing have re-introduced multi-chip module (MCM) architectures as an alternative to the current monolithic approach. MCMs or multi-die systems implement multiple smaller chiplets in a single package. These MCMs are connected through various package interconnect technologies, such as current industry solutions in AMD's Infinity Fabric, Intel's Foveros active interposer, and Marvell's Mochi Interconnect. Although MCMs improve manufacturing yields and are cost-effective, additional challenges on the Network-on-Chip (NoC) within a single chiplet and across multiple chiplets need to be addressed. These challenges include routing, scalability performance, and resource allocation. This work introduces a scalable MCM 3D interconnect infrastructure called ''MCM-3D-NoC'' with multiple 3D chiplets connected through an active interposer. System-level simulations of MCM-3D-NoC are performed to validate the proposed architecture and provide performance evaluation of network latency, throughput, and EDP.

CCS CONCEPTS

•Computer systems organization → Integrated Circuits; *VLSI (very large scale integration);*

KEYWORDS

3D NoCs, Interconnect Network, Multi-Die Systems, Multi-Chip Modules

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.
NOCS '19, New York, NY, USA
© 2019 ACM. 978-1-4503-6700-4/19/10…$15.00
DOI: 10.1145/3313231.3352380

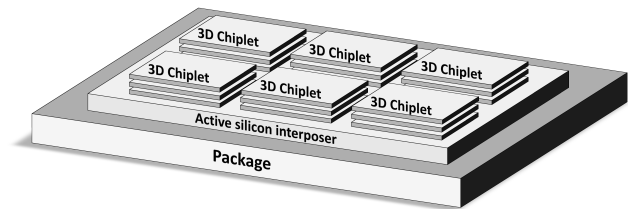


Figure 1: MCM with active interposer and 3D chiplets.

1 INTRODUCTION

Monolithic fabrication is the predominantly preferred method for the manufacturing of contemporary computing systems. All major components of the system, including I/O, memory, CPU cores, and GPU cores, are all integrated within a single die on a single silicon technology. The major detriments of these monolithic designs include increased costs, reduced fabrication yields, and limited scalability [1]. Therefore, multi-chip modules (MCMs), which break the monolithic structure into multiple smaller, higher yielding die, are actively researched in industry and academia [2–5].

Although multi-chip modules have been explored in the past, they did not gain traction in industry due to the limited package interconnect technology for inter-die communication. Current MCMs implement different package interconnect approaches which include AMD's server-grade EPYC and Threadripper processors [3], Intel's Foveros and Embedded Multi-die Interconnect Bridge (EMIB) [4], active interposer research [5], and NVIDIA's research in multi-chiplet with MCM-GPU [2].

Although higher yield is achieved with the introduction of multi-die systems, some performance degradation is expected due to the decrease in connectivity of the processing elements (PEs) across multiple die. The decrease in connectivity of distant PEs especially evident in the case of PEs distributed over multiple 3D die (i.e 3D chiplets with a 3D NoC). This exemplary topology is depicted in Figure 1, with six 3D chiplets connected through the active silicon interposer over the package. The 3D chiplet topology with a 3D

NoC is the topology of choice for the recently introduced industrial solution for Intel's Foveros [6].

In order for a packet to reach a destination in a different 3D chiplet, it has to first traverse the layers in the local 3D chiplet, traverse the active silicon interposer layer, and then reach the PE at a layer of the destination 3D chiplet. Similar performance degradation (although to a lesser extent) can been seen in 2D ICs with 2D NoCs across multiple die [1, 5].

In this work, a MCM-3D-NoC is proposed to provide a scalable solution for future generation processors. MCM-3D-NoC connects multiple 3D chiplets on a single package through an active interposer, which acts as an interconnect layer for all the chiplets. An illustrative package with six 3D chiplets connected with an active interposer is shown in Figure 1. Each chiplet has three layer and the active silicon acts as an interconnect layer with dedicated logic to connect the chiplets in the package. Although the illustration shows an homogeneous architecture, MCMs are capable of heterogeneous designs at a lower cost and higher yield than monolithic fabrication. The benefits of MCMs vs. monolithic, which include scalability, modular design, and higher yield, are largely unquantifiable, although system-level evaluation provides useful insights on the performance of future MCM-enabled systems. Proposed MCM-3D-NoC encapsulates a custom inter-chiplet routing algorithm for improved latency and throughput. Additionally, resource allocation and scalability analysis are performed to evaluate the impact of various chiplet configurations. MCM-3D-NoC system performance is evaluated with a custom SystemC cycle-accurate network simulator.

A background on current package interconnect technologies and related works are presented in Section 2. The proposed MCM-3D-NoC is detailed in Section 3. The custom inter-chiplet routing algorithm is discussed in Section 4. Simulation setup and performance evaluation of MCM-3D-NoC are presented in Section 6. Finally, the conclusions of this work are presented in Section 7.

2 BACKGROUND AND RELATED WORKS

Multi-chip modules are implemented in consumer-level products [3] to tackle scalability challenges that arise from manufacturing a monolithic single-chip multi-IP package. These multi-chiplet packages improve manufacturing yields considerably [1] and improve processing scalability for future-generation workloads targeting exascale computation [7]. Current die-to-die connectivity includes Intel's Foveros and EMIB [4], AMD's Infinity Band [3], Marvell's Mochi interconnect [8], and passive/active silicon interposers [1, 5]. Each of these interconnects has their respective detriments and benefits, measured primarily in 1) bandwidth, 2) power/energy profile, and, 3) ease of scalability.

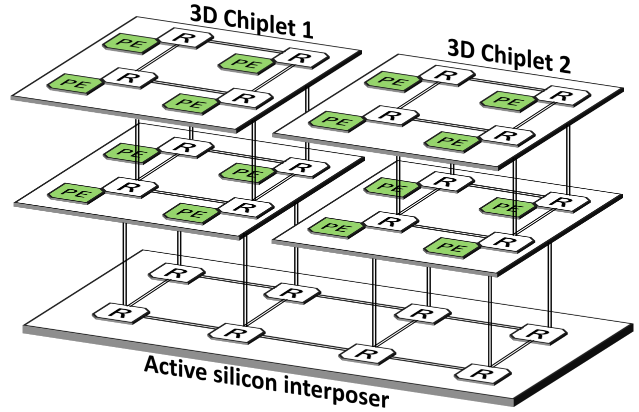
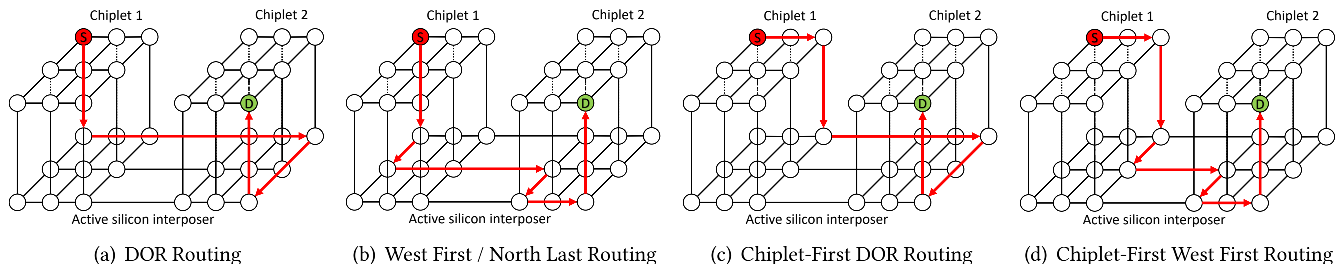


Figure 2: MCM-3D-NoC architecture scaled down for clarity. Each 3D chiplet can represent IP blocks, memory, or compute units.

Jerger et al. [1, 5] and Vivet et al. [9] investigate the feasibility of active silicon interposers for 3D ICs integration, including heterogeneous systems. Stow et al. [10] and Coskun et al. [11] perform a cost and performance comparison between traditional monolithic 2D SoCs, 2.5D passive interposers, and 2.5D/3D active interposers to demonstrate the trade-offs between the interposer types for current and future high-performance systems. Yin et al. [12] introduce a modular deadlock-free routing algorithm for 2D NoCs on chiplet-based systems with active interposers. Kuttappa et al. [13] propose a novel clock generation and distribution network for multi-die architectures connected through an active silicon interposer. The proposed clock network generates and distributes a resonant clock through the active silicon interposer between dies, with each die served through resonant local clock trees.

One major thrust in relevant research is towards, photonic NoCs, which have gained interest to address bandwidth challenges associated with many core systems [14, 15]. Abellan et al. [14] study the bandwidth challenges associated with silicon-photon links with different NoC architectures for 2D-mesh based designs. Narayan et al. [15] propose a wavelength selection technique along with a framework to model system performance and power for 2.5D systems.

Research on monolithic (as opposed to MCM) 3D NoCs, include Pavlidis et al. [16] and Feero et al [17], which provide performance evaluation of NoCs implemented on 3D ICs to demonstrate the superior functionality in terms of throughput and latency compared to traditional 2D NoCs. Vivet et al. [18] have fabricated a 4×4×2 3D NoC prototype on an active interposer. The versatility of the 3D NoC topologies with arbitrary 2D topologies in each of the individual layers and partially connected layers is shown in [19]. The elevator placement is optimized in [20] using a heuristic, significantly improving the 3D NoC performance. More et



(a) DOR Routing (b) West First / North Last Routing (c) Chiplet-First DOR Routing (d) Chiplet-First West First Routing

Figure 3: The different routing algorithms evaluated for `MCM-3D-NoCs`.

al. [21] investigate an arbitration-free design for the shared vertical channels and compared with other 3-D NoC architectures using traditional synthetic traffic patterns and Rentian traffic emulating applications for CMPs.

To the best of the author's knowledge, this work is the first to introduce and evaluate a non-monolithic, multi-chiplet, 3D NoC.

3 MULTI-CHIP MODULE 3D NOC

A smaller representative model of the proposed `MCM-3D-NoC` with active interposer is depicted in Figure 2. The base layer of the topology is located on the shared active silicon interposer, which is capable of additional logic, as opposed to a passive silicon interposer. The active interposer does not have any complex processing elements, memory, or additional IPs; the base layer acts only as a routing infrastructure for the 3D chiplets. The upper layers represent the 3D chiplets, which are not connected to each-other except for the base layer on the active interposer. These 3D chiplets are modular, and additional chiplets of various dimensions and purposes can be added to `MCM-3D-NoC` easily.

For the purpose of clarity, the evaluated `MCM-3D-NoC` topology in this work is homogeneous and arbitrarily defined as 9 chiplets, with 3 layers each, and 4×4 routers per layer. The 3D chiplets are positioned in a 3×3 grid. Although, scalability analysis performed in Section 6.3, evaluates a separate larger topology for thoroughness.

Concerted effort is made to keep the router architecture relatively simplistic to avoid overly complicated logic on the active silicon interposer. Nonetheless, resource allocation analysis performed in Section 6.2 evaluates the impact of additional buffer configurations on the `MCM-3D-NoC`. The router is equipped with 4 buffers per input port and 2 virtual channels to guarantee deadlock and livelock avoidance.

There are a maximum of 7 ports in the 3D chiplet routers, 4 for the cardinal directions, 2 for the up and down directions, and an additional for local traffic (i.e. PEs, IPs, memory). The routers on the base layer do not have additional ports for local traffic and for the down direction, therefore conserving energy and decreasing area overhead. The base layer has a one-to-one router pairing with each of the 3D chiplets, therefore the base layer has a total of 12×12 routers. Any transaction between 3D chiplets needs to traverse the base layer on the active silicon interposer.

4 ROUTING ALGORITHM

The routing algorithm used for `MCM-3D-NoC` considers the source chiplet, destination chiplet, and active interposer layer. The interposer layer does not generate or receive traffic (as shown in Figure 2), therefore all the traffic generated in the network has a source and destination pair located in any of the 3D chiplets present in system. If both the source and destination chiplets are the same then the packet is routed using deterministic XYZ routing algorithm.

If the destination chiplet is different from the source then the packet is first routed to the base layer for intra-die communication. On the base layer, the packet is routed to the destination XY coordinates and then to the Z layer of the destination chiplet.

Further details regarding the different routing algorithms evaluated, including partially-adaptive *West First* and *North Last*, are included in Sections 4.1–4.3.

4.1 Baseline Algorithm

In order to route a packet with a destination to a different 3D chiplet and currently located on the base layer, the baseline algorithm utilizes a deterministic dimension order routing (DOR) XY algorithm to reach the XY destination coordinates and then route to the destination Z layer.

Two partially-adaptive algorithms for the base layer, West First and North Last, are implemented to quantify the performance benefit of the increase network utilization that partially-adaptive algorithms provide. A fully-adaptive routing algorithm is not implemented due to the hardware constrain imposed by the fabrication of logic on silicon interposer. The baseline routing algorithms are shown in Figure 3(a)–3(b). Both partially-adaptive algorithms can pick between different shortest path routes based on the buffer availability at each intermediate node, as shown in Figure 3(b).

4.2 Chiplet-First Algorithm

The "Chiplet-First" routing algorithm is shown in Figure 3(c). The purpose of the "Chiplet-First" algorithm is to improve network utilization on the 3D chiplet and decrease network congestion on the base layer by bringing the packet closer to the destination before entering the active silicon interposer.

If a packet destination is located on a different 3D chiplet, and the source node is located on the *top-most* layer of the 3D chiplet then the packet is routed to the edge of the 3D chiplet first then routed the base layer. The packet is routed towards a specific edge based on the XY distance of the destination node. If the destination node is further in the X direction then the packet is routed to the left or right edge of the 3D chiplet. If the destination is equidistant from the source node then the edge is selected based on the buffer availability at the intermediate node.

4.3 Chiplet-First West First Algorithm

The Chiplet-First West First algorithm combines both the previously detailed Chiplet-First algorithm and the partially-adaptive West First algorithm on the base layer. The Chiplet-First West First routing algorithm is shown in Figure 3(d). The deterministic XYZ routing algorithm used for source-destination pairs within a 3D chiplet remains the same across all different routing algorithms. This decision is made to accurately quantify the performance improvement to the base layer only, either directly by implementing partially-adaptive routing algorithms, or indirectly by implementing Chiplet-First routing algorithm which decreases network congestion on the base layer.

5 DECONSTRUCTING 3D NOCS FOR MULTI-DIE SYSTEMS

Monolithic 3D NoCs have the benefit of added connectivity between the layers as opposed to transferring packets to the base layer for inter-die communication. This addition in connectivity and performance improvement in monolithic 3D NoCs are more evident with each added IC layer. The primary challenge that next generation systems have to solve is the communication bottleneck between vertical and horizontal transmissions. For each added layer in a 3D IC, the base transmission layer between multiple die is one hop further. Bus-based communication techniques like the Elevator-First [19] or arbitration-free vertical channels [21] have partially solved long-distance vertical communication but no studies have implemented these algorithms and architectures for multi-die systems. In addition, the limited resources on the base layer increase network congestion and delay across the entire system as opposed to only a particular die with heavy load.

In order to fully evaluate the impact of chiplet-based systems in 3D IC, performance comparison between a monolithic 3D NoC and MCM 3D NoC is detailed in Figures 4-5. A cycle-accurate SystemC network simulator is used to perform `MCM-3D-NoC` performance evaluation. The base layer mesh size arbitrarily chosen for this experimental evaluation is 12×12 (total 144 routers). On the active interposer, are nine 3D chiplets on a 3×3 grid with dimensions of 4×4×3. The monolithic 3D NoC is in turn sized to a 12×12×4 to preserve the same structure and number of routers in the network. Additionally, networks with 2 and 3 number of layers are evaluated to quantify the performance change. The non-adaptive XYZ routing algorithm is used to route packets for the monolithic 3D NoC and the baseline algorithm discussed in Section 4.1 is used to route packets for MCM 3D NoC.

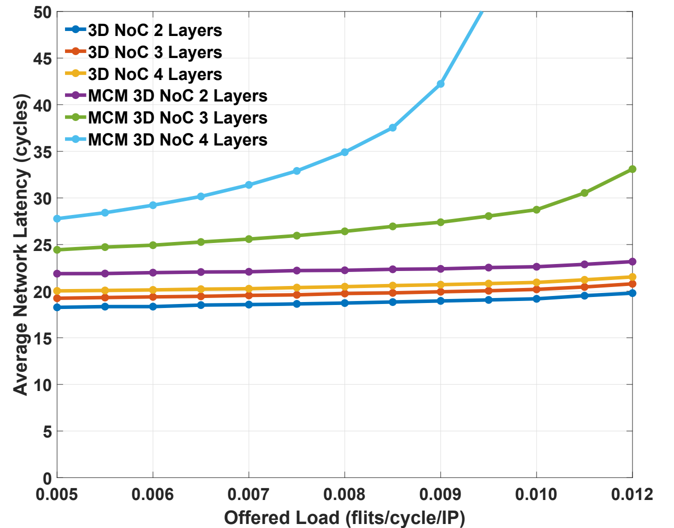


Figure 4: Latency comparison of monolithic 3D NoC with `MCM-3D-NoC` under uniform random traffic.

Latency: Average network latency across multiple flit injection rates is shown in Figure 4. With the increase of flit injection rate, latency of MCM 3D NoC increases considerably faster compared to monolithic 3D NoC. There is a clear divide between latency in monolith 3D NoC and MCM-3D-NoC, particularly with the increase of the number of layers of IC. This increase in latency is attributed to increase in network traffic to the base layer from the additional layers of the 3D IC, essentially creating a bottleneck in the inter-die communication. When the 3D NoC is provisioned with 2 layers only, the difference in latency is negligible as seen also in [12]. The increase in latency is not seen at all with the monolithic 3D NoC due to the highly connected structure of the topology which doesn't require traversal to the base layer.

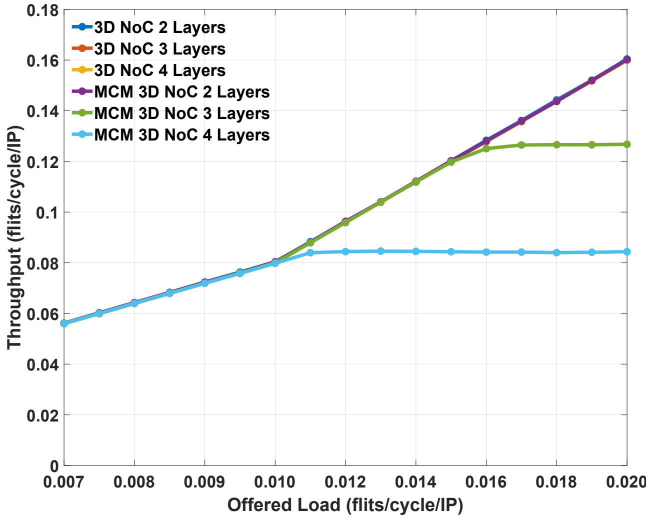


Figure 5: Throughput comparison of monolithic 3D NoC with `MCM-3D-NoC` under uniform random traffic.

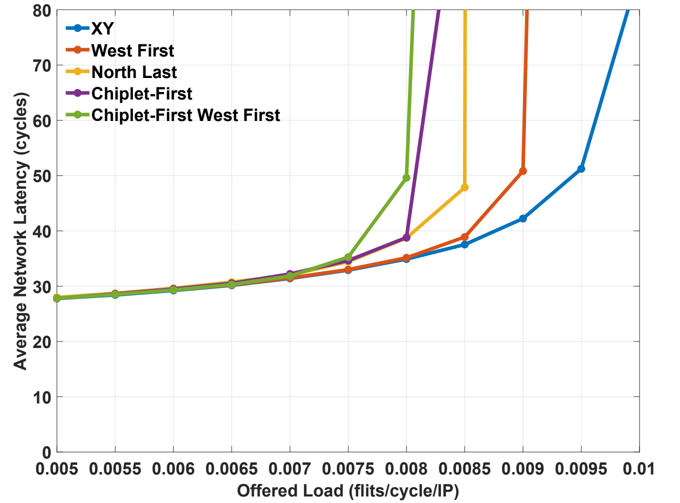


Figure 6: `MCM-3D-NoC` average latency under uniform random traffic with various routing algorithms.

Throughput: Network throughput across multiple flit injection rates is shown in Figure 5. Similar behavior to latency, network throughput saturates faster and lower when separating the monolithic 3D NoC into smaller 3D chiplets in the MCM-3D-NoC architecture. Even though the 2 layer topology has particularly good throughput (as compared to the monolithic 3D NoC), the increase in number of layers quickly decreases throughput and saturates the network at lower flit injection rates. The monolithic 3D NoC is not phased by the increase of layers as packet traversal does not require sidetracking to the base layer, therefore increasing overall throughput of the network.

Although the performance benefits of monolithic 3D NoC cannot be ignored, successful fabrication of a large monolithic 3D NoC would require drastic changes to the manufacturing process. Primarily to improve yield and decrease fabrication costs, industry has moved to a chiplet-based manufacturing process which would in turn permit robust fabrication of large scale computing packages capable of hundreds of PEs.

6 MCM-3D-NOC PERFORMANCE EVALUATION

Power estimations are performed utilizing the latest version of DSENT [22]. The packet size is 8 flits and the flit size is 32 bits. Five traffic patterns are used to analyze `MCM-3D-NoC`: 1) uniform random, 2) 25% localized (to 3D chiplet) traffic, 3) 50% localized traffic, 4) 75% localized traffic, and 5) hotspot. The performance metrics evaluated are: 1) throughput, 2) average latency, and 3) EDP per flit.

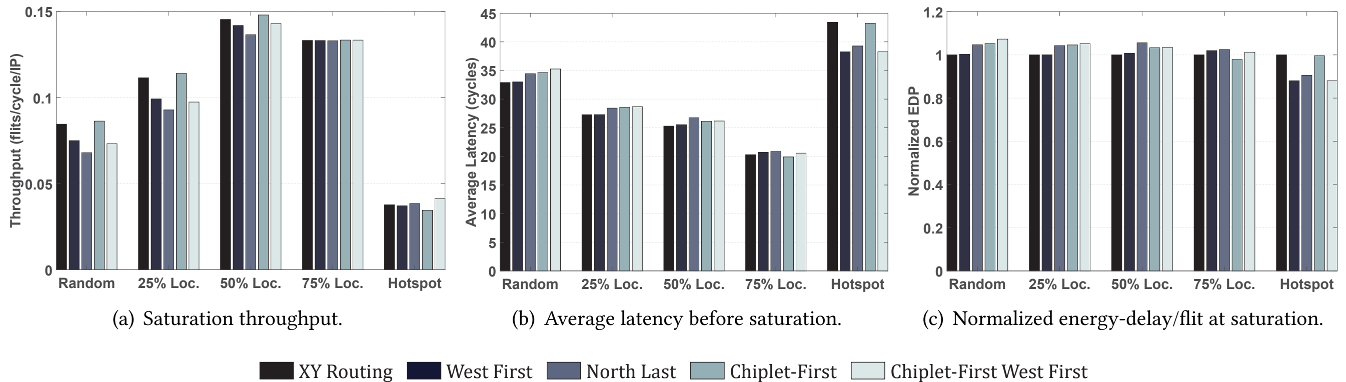
Initially, the average network latency is swept across multiple packet injection rates to evaluate both low-load latency and saturation throughput. Later, performance is evaluated at network saturation. The average network latencies of `MCM-3D-NoC` with various routing algorithms under *uniform random* traffic, are shown in Figure 6.

The results indicate that for *uniform random* traffic, the non-adaptive XY routing algorithm outperforms all other partially-adaptive algorithms at high traffic load. The XY routing algorithm has more global, long-term information about the characteristics of uniform traffic, which leads to even distribution of traffic across the `MCM-3D-NoC`. The Chiplet-First routing algorithm performs poorly under *uniform random* traffic, due to also being partially-adaptive and routing to the edges of the 3D chiplet may cause congestion at the top layer, therefore increasing latency. This latency evaluation has similar results compared to the 2D NoC with active silicon interposer latency evaluation performed in [12].

6.1 Simulation Results

Simulations results of `MCM-3D-NoC` with various routing algorithms under different traffic are shown in Figure 7.

Throughput: Throughput results are shown in Figure 7(a). The maximum throughput achieved under *uniform random* traffic is with Chiplet-First routing algorithm, at 0.086 flits/cycle/IP. Partially-adaptive algorithm, including Chiplet-First West First, show a decrease in network throughput with *uniform random* traffic of ≈20% compared to XY routing. Deterministic routing algorithms perform particularly well under uniformly distributed traffic. `MCM-3D-NoC` under *localized* traffic pattern performs similarly to the *uniform random* traffic. In the *hotspot* traffic pattern simulated, 4 close nodes



(a) Saturation throughput. (b) Average latency before saturation. (c) Normalized energy-delay/flit at saturation.

Figure 7: Simulation results of `MCM-3D-NoC` for varying traffic patterns and routing algorithms.

are designated as the hot spot nodes, which receive traffic in addition to the regular uniform traffic. The nodes are located at the center 3D chiplet on the top-most layer. The network under *hotspot* traffic pattern saturates sooner and throughput is not as high as the rest of the traffic pattern. It can be seen that partially-adaptive routing algorithms perform better than the XY algorithm, due to the non-uniformity of the traffic pattern.

Latency: Network latency evaluation is shown in Figure 7(b). As shown in Figure 6, XY routing algorithm is the preferred choice for `MCM-3D-NoC` running *uniform random* traffic. Not only the network saturates sooner under partially-adaptive algorithms but also the latency is higher when not saturated. But when `MCM-3D-NoC` runs non-uniform traffic like the *hotspot* traffic pattern, the partially-adaptive algorithms perform significantly better than XY routing, up to ≈12% with Chiplet-First West First routing algorithm.

Energy-delay product: Energy-delay product evaluation is shown in Figure 7(c). Energy estimations are performed using the latest version of DSENT [22] using the 22nm technology node. EDP is useful because network latency and energy are both taken into account, therefore characterizing the improvement in performance considering the added energy consumption. `MCM-3D-NoC` running *hotspot* traffic pattern displays an improvement in EDP up to ≈12%. In the rest of the traffic patterns analyzed, EDP is within ≈5% of the XY routing algorithm.

6.2 Resource Allocation Analysis

Simulations results of `MCM-3D-NoC` with additional resources under different traffic patterns are shown in Figure 8. The `MCM-3D-NoC` is evaluated with increased buffer depth across the entire system and only the routers on the base layer (denoted by *"Extra"*). The standard deterministic XY routing algorithm is used to route packets across the network. The first two test cases demonstrate the impact of additional buffer space on the base layer compared to the previously evaluated `MCM-3D-NoC`.

The second set of cases, demonstrate the impact of distributing resources equally across the 3D chiplets and the base layer on the active silicon interposer, or dedicating the extra resources only for the base layer. The same number of buffers are used on `MCM-3D-NoC`, only their distribution differs (6 buffers across vs. 4 buffers across and 8 extra on the routers of the base layer).

Throughput: Throughput results are shown in Figure 8(a). With the addition of the 4 extra buffers on each router on the base layer, throughput is improved up to ≈10% in *uniform random* and *localized* traffic patterns. `MCM-3D-NoC` under *hotspot* traffic pattern shows minimal change with the addition of buffers on the base layer.

Allocating the additional resources to the routers on the base layer as opposed as throughout the 3D chiplets, improves throughput up to ≈10%. The base layer acts as an interconnect for the many 3D chiplets and any additional resource helps improve the performance of the network. Therefore, at the same cost of area and buffer energy consumption, it is advised to dedicate additional resource to the base layer as opposed to all system components.

Latency: Network latency evaluation is shown in Figure 8(b). The addition of buffer space has a minimal impact on latency across all traffic patterns and configurations. Although, latency improves ≈8% on `MCM-3D-NoC` under *hotspot* traffic pattern if additional resources are dedicated on the base layer.

Energy-delay product: Energy-delay product evaluation is shown in Figure 8(c). With the addition of buffer space, energy consumption is consequently increased. Therefore EDP is increased across all traffic patterns and configurations. It is worth noting that under *hotspot* traffic pattern, EDP is increased only ≈20% if resources are dedicated to the base layer, as opposed to ≈31% if shared across the system.

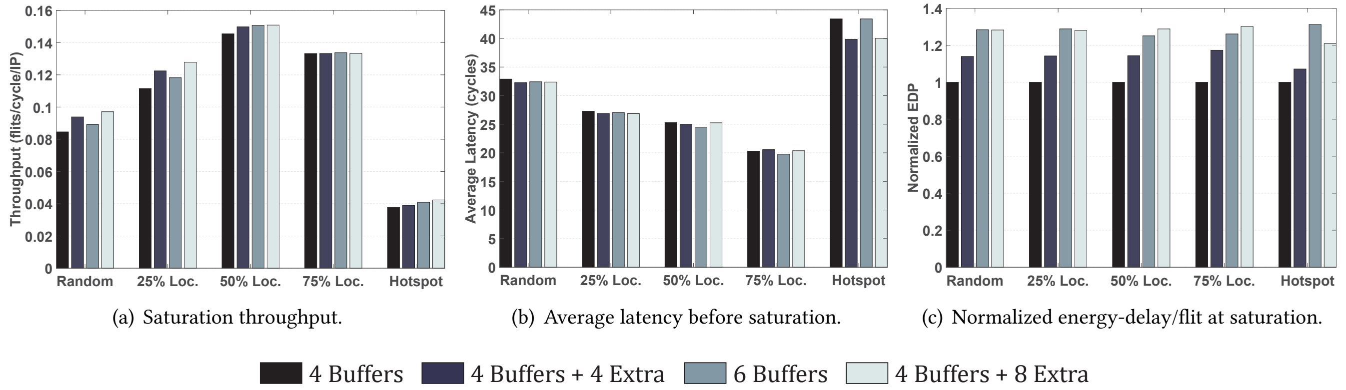


Figure 8: Simulation results with increased resources for varying traffic patterns and XY routing algorithm.

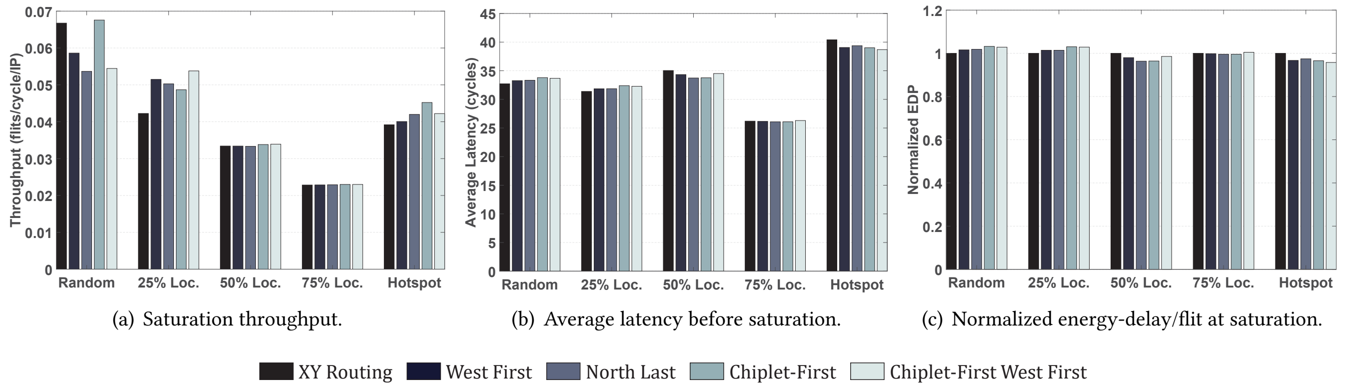


Figure 9: Simulation results of larger `MCM-3D-NoC` for varying traffic patterns and routing algorithms.

6.3 Scalability Analysis

A larger 4×16×16 `MCM-3D-NoC` with 4 3D chiplets on a 2×2 grid is simulated for scalability analysis. Simulations results of the larger `MCM-3D-NoC` with various routing algorithms under different traffic patterns are shown in Figure 9. The evaluation investigates the performance impact of having fewer and bigger 3D chiplets on the MCM as opposed to the previous evaluation of 9 3D chiplets.

Throughput: Throughput results are shown in Figure 9(a). Similar to the smaller `MCM-3D-NoC`, under *uniform random* traffic, throughput is higher with XY routing algorithm. `MCM-3D-NoC` under 25% *localized* traffic pattern, shows an increase in throughput when utilizing partially-adaptive routing algorithms, achieving up to ≈27% throughput improvement with the Chiplet-First West First routing algorithm.

This improvement is attributed to the larger 3D chiplet size and fewer 3D chiplets on the network, therefore lowering congestion and improving throughput. The MCM under other traffic has similar throughput across routing algorithms. The 4 nodes that comprise the hotspot which resides on the center of the network is split across one corner of each 3D chiplets, therefore `MCM-3D-NoC` using the Chiplet-First routing algorithm has improved throughput of ≈15%.

Latency: Network latency evaluation is shown in Figure 9(b). Under *uniform random* traffic, latency remains within ≈3%, across all routing algorithms. An improvement in latency of ≈4% can be seen on `MCM-3D-NoC` under 50% *localized* and *hotspot* traffic pattern.

Energy-delay product: Energy-delay product evaluation is shown in Figure 9(c). Due to the similarity in latency, EDP variance is within ≈4% across all traffic patterns and routing algorithms. `MCM-3D-NoC` under *hotspot* traffic pattern performs best with the Chiplet-First West First routing algorithm.

7 CONCLUSIONS

This work proposes a network interconnect for multi-die systems capable of linking multiple 3D chiplets with an active

silicon interposer. The active interposer acts as an interconnect layer that allows for modular plug-and-play heterogeneous exascale architectures. To verify functionality and performance, a SystemC `MCM-3D-NoC` simulator is evaluated under various traffic patterns and routing algorithms.

A custom "Chiplet-First" routing algorithm is implemented which improves network congestion on the base interconnect layer. Multiple routing algorithms are evaluated and it is found that non-uniform traffic patterns have an improvement in throughput up to ≈15%. In addition, resource allocation analysis has concluded that over-provisioning the base layer, improves throughput, latency and EDP up to ≈10%. Scalability analysis indicates that `MCM-3D-NoC` can be adapted to multiple network configurations and 3D chiplet sizes with minimal impact on performance.

The benefits of MCM design, which include enhanced scalability, modularity, and higher fabrication yields, go beyond system-level performance metrics. In that regard, `MCM-3D-NoC` is developed to provide a thorough analysis and evaluation of challenges and solutions in the MCM interconnect network.

REFERENCES

[1] A. Kannan, N. E. Jerger, and G. H. Loh, "Enabling Interposer-based Disintegration of Multi-core Processors," in *Proceedings of the IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pp. 546–558, December 2015.

[2] A. Arunkumar, E. Bolotin, B. Cho, U. Milic, E. Ebrahimi, O. Villa, A. Jaleel, C. Wu, and D. Nellans, "MCM-GPU: Multi-chip-module GPUs for Continued Performance Scalability," in *Proceedings of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, pp. 320–332, June 2017.

[3] N. Beck, S. White, M. Paraschou, and S. Naffziger, "'Zeppelin': An SoC for Multichip Architectures," in *Proceedings of the IEEE International Solid State Circuits Conference (ISSCC)*, pp. 40–42, February 2018.

[4] R. Mahajan *et al.*, "Embedded Multi-die Interconnect Bridge (EMIB) – A High Density, High Bandwidth Packaging Interconnect," in *Proceedings of IEEE Electronic Components and Technology Conference (ECTC)*, pp. 557–565, May 2016.

[5] N. E. Jerger, A. Kannan, Z. Li, and G. H. Loh, "NoC Architectures for Silicon Interposer Systems," in *Proceedings of the IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pp. 458–470, December 2014.

[6] Intel Foveros Interconnect. https://spectrum.ieee.org/tech-talk/semiconductors/processors/intel-shows-off-chip-packaging-powers, 2019.

[7] U.S. Department of Energy. https://www.exascaleproject.org/, 2018.

[8] Marvell Corp. https://www.marvell.com/architecture/mochi/, 2018.

[9] P. Vivet, C. Bernard, F. Clermidy, D. Dutoit, E. Guthmuller, I. Panadès, G. Pillonnet, Y. Thonnart, A. Garnier, D. Lattard, A. Jouve, F. Bana, T. Mourier, and S. Chéramy, "3D Advanced Integration Technology for Heterogeneous Systems," in *Proceedings of the International 3D Systems Integration Conference (3DIC)*, August 2015.

[10] D. Stow, Y. Xie, T. Siddiqua, and G. H. Loh, "Cost-effective Design of Scalable High-performance Systems using Active and Passive Interposers," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pp. 728–735, November 2017.

[11] A. Coskun, F. Eris, A. Joshi, A. B. Kahng, Y. Ma, and V. Srinivas, "A cross-layer methodology for design and optimization of networks in 2.5d systems," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pp. 1–8, November 2018.

[12] J. Yin, Z. Lin, O. Kayiran, M. Poremba, M. S. B. Altaf, N. E. Jerger, and G. H. Loh, "Modular Routing Design for Chiplet-Based Systems," in *Proceedings of the ACM/IEEE International Symposium on Computer Architecture (ISCA)*, pp. 726–738, June 2018.

[13] R. Kuttappa, B. Taskin, S. Lerner, V. Pano, and I. Savidis, "Robust low power clock synchronization for multi-die systems," in *Proceedings of the ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED)*, pp. 1–6, July 2019.

[14] J. L. Abellán, C. Chen, and A. Joshi, "Electro-photonic noc designs for kilocore systems," *Journal of Emerging Technologies in Computing Systems*, vol. 13, pp. 24:1–24:25, November 2016.

[15] A. Narayan, Y. Thonnart, P. Vivet, C. F. Tortolero, and A. K. Coskun, "WAVES: Wavelength selection for power-efficient 2.5d-integrated photonic nocs," in *Proceedings of the ACM/IEEE Design, Automation Test in Europe Conference Exhibition (DATE)*, pp. 516–521, March 2019.

[16] V. F. Pavlidis and E. G. Friedman, "3-D Topologies for Networks-on-Chip," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 15, pp. 1081–1090, October 2007.

[17] B. S. Feero and P. P. Pande, "Networks-on-Chip in a Three-Dimensional Environment: A Performance Evaluation," *IEEE Transactions on Computers*, vol. 58, pp. 32–45, January 2009.

[18] P. Vivet, Y. Thonnart, R. Lemaire, C. Santos, E. Beigné, C. Bernard, F. Darve, D. Lattard, I. Miro-Panadès, D. Dutoit, F. Clermidy, S. Cheramy, A. Sheibanyrad, F. Pétrot, E. Flamand, J. Michailos, A. Arriordaz, L. Wang, and J. Schloeffel, "A 4x4x2 Homogeneous Scalable 3D Network-on-Chip Circuit With 326 MFlit/s 0.66 pJ/b Robust and Fault Tolerant Asynchronous 3D Links," *IEEE Journal of Solid-State Circuits*, vol. 52, pp. 33–49, January 2017.

[19] F. Dubois, A. Sheibanyrad, F. Petrot, and M. Bahmani, "Elevator-first: A deadlock-free distributed routing algorithm for vertically partially connected 3d-nocs," *IEEE Transactions on Computers*, vol. 62, no. 3, pp. 609–615, 2013.

[20] S. Foroutan, A. Sheibanyrad, and F. Petrot, "Assignment of vertical-links to routers in vertically-partially-connected 3-d-nocs," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, vol. 33, no. 8, pp. 1208–1218, 2014.

[21] A. More, V. Pano, and B. Taskin, "Vertical arbitration-free 3D NoCs," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, pp. 1–1, October 2017.

[22] C. Sun, C. O. Chen, G. Kurian, L. Wei, J. Miller, A. Agarwal, L. Peh, and V. Stojanovic, "DSENT - A Tool Connecting Emerging Photonics with Electronics for Opto-Electronic Networks-on-Chip Modeling," in *Proceedings of the IEEE/ACM International Symposium on Networks-on-Chip (NoCs)*, pp. 201–210, May 2012.