Thesis for The Degree of Licentiate of Engineering

Design, Analysis, and Modeling of On-chip Interconnects for Large Scale Systems

Panagiotis Strikos

Division of Computer and Network Systems
Department of Computer Science & Engineering
Chalmers University of Technology
Gothenburg, Sweden, 2026

Design, Analysis, and Modeling of On-chip Interconnects for Large Scale Systems

Panagiotis Strikos

Advisor:
Professor Ioannis Sourdis, Chalmers University of Technology

Co-Advisors:
Ahsen Ejaz, PhD, InfiniNode Technologies

Examiner:
Professor Pedro Petersen Moura Trancoso, Chalmers University of Technology

Discussion Leader:
Georgios Michelogiannakis, PhD, Microsoft

Copyright ©2026 Panagiotis Strikos
except where otherwise stated.
All rights reserved.

Department of Computer Science & Engineering
Division of Computer and Network Systems
Chalmers University of Technology and Gothenburg University
Gothenburg, Sweden

This thesis has been prepared using LaTeX.
Printed by Chalmers Reproservice,
Gothenburg, Sweden 2026.

It is what it is...

Abstract

As the semiconductor industry faces increasing challenges in technology scaling, integrading more devices onto a single monolithic die incurs prohibitive manufacturing costs due to low yield. Chiplet-based systems have emerged as a cost-effective alternative, offering higher yield and enabling heterogeneous integration across different process technologies. However, chiplet-based systems come with performance overheads. In particular, inter-chiplet communication is constrained by limited bandwidth and higher latency overheads that are further exacerbated by the use of passive silicon interposers, which do not include any active components and limit the network design. Studying these overheads requires an infrastructure capable of simulating such systems. To that end, this thesis produces BZSim, an open-source framework for fast, large-scale microarchitectural simulation with detailed interconnect modeling. BZSim integrates BookSim into ZSim, combining accurate network modeling with fast parallel execution. It detects low-contention network traffic and analytically calculates its latency instead of simulating it, reducing the simulation overhead. This way, it achieves 2–4× faster simulations with a maximum normalized IPC error of 3–5% and an average packet latency error of 10–20%. Therefore it is an order of magnitude faster than gem5 while remaining just 3× slower than standalone ZSim. While the cost and yield advantages of chiplet-based designs are well established, their system performance overheads relative to monolithic designs have not been systematically studied. Using BZSim, this thesis conducts a comprehensive analysis of chiplet-based systems across a range of design parameters, measuring performance overheads relative to monolithic designs and their scalability to system and chiplet size. The study covers design alternatives in memory hierarchy organization, interconnect choices, and technology aspects. The analysis reveals that chiplet-based chips reduce recurring engineering costs by almost half compared to monolithic designs, at the cost of about a third of the monolithic performance. The inter-chiplet interconnect identified as a key source of this overhead, motivating a targeted network solution. For this reason, this thesis proposes PRISM: a new network for multi-chiplet chips on passive silicon interposers. PRISM introduces a new topology which offeres low hop count for inter-chiplet traffic, and a deterministic routing algorithm with virtual-network separation for deadlock-free routing. It also improves chiplet-to-chiplet bandwidth utilization through selective flit compression, applied dynamically per flit only when inter-chiplet congestion is detected, minimizing compression overheads. Evaluated on a 256-core system with 16 chiplets, PRISM improves inter-chiplet packet latency by up to 5.8× and overall system performance by up to 17.4%, with selective compression outperforming always-on compression by up to 6.5%.

Keywords: Chiplet-based systems, Network-on-Chip, Microarchitectural simulation, In-network compression

Acknowledgments

First and foremost, I would like to thank my supervisor Yiannis Sourdis and co-supervisor Ahsen Ejaz for their continuous guidance, support, and invaluable feedback throughout this work. My deepest gratitude goes to my officemates Konstantinos, Magnus, and especially Neethu, with whom I had the pleasure of collaborating. The discussions, shared ideas, and camaraderie you all brought to everyday academic life made this journey far more enjoyable and meaningful. I would also like to thank all my other current and former colleagues for the stimulating and friendly environment they helped create.

Finally, to my family and friends: thank you for always being there for me.
Σας ευχαριστώ για όλα.

This work is supported by the Swedish Foundation for Strategic Research (contract number CHI19-0048) under the PRIDE project.

List of Publications

Appended publications

This thesis is based on the following publications:

[I] **Panagiotis Strikos**, Ahsen Ejaz, and Ioannis Sourdis
BZSim: Fast, Large-scale Microarchitectural Simulation with Detailed Interconnect Modeling
2024 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS).

[II] Neethu Bal Mallya, **Panagiotis Strikos**, Bhavishya Goel, Ahsen Ejaz, and Ioannis Sourdis
A Performance Analysis of Chiplet-Based Systems
2025 Design, Automation and Test in Europe Conference and Exhibition (DATE), Lyon, France, Mar 2025.

[III] **Panagiotis Strikos**, Ahsen Ejaz, and Ioannis Sourdis
PRISM: Reducing Performance Overhead of Multi-chiplet Chip Interconnects
Under Submission.

Contents

Abstract v

Acknowledgments vii

List of Publications ix

1 Introduction 1
- 1.1 Problem Statement . . . 2
  - 1.1.1 The Simulation Scalability Challenge . . . 2
  - 1.1.2 Performance Analysis of Chiplet-Based Systems . . . 3
  - 1.1.3 Reducing Inter-Chiplet Communication Overhead . . . 3
- 1.2 Thesis Objectives and Contributions . . . 3
  - 1.2.1 Fast, Large-Scale Simulation with Detailed Interconnect Modeling . . . 3
  - 1.2.2 Performance Analysis of Chiplet-Based Systems . . . 5
  - 1.2.3 Reducing Inter-Chiplet Communication Overhead . . . 6
- 1.3 Thesis Outline . . . 8

Bibliography 9

I Included Papers 13

2 BZSim: Fast, Large-scale Microarchitectural Simulation 15
- 2.1 Introduction . . . 17
- 2.2 Background . . . 19
  - 2.2.1 ZSim and other microarchitectural simulators . . . 20
  - 2.2.2 BookSim and other network simulators . . . 20
  - 2.2.3 Current state-of-the-art high-performance NoCs . . . 21
- 2.3 BZSim: exposing the tradeoff between speed and accuracy of network modeling . . . 23
  - 2.3.1 Detecting and skipping simulation of low contention traffic 23
  - 2.3.2 Skipping the evaluation of empty network parts . . . 25
  - 2.3.3 BookSim-ZSim integration . . . 25
- 2.4 Modeling of high performance NoCs . . . 25
- 2.5 Experimental setup . . . 28
- 2.6 Evaluation . . . 29
  - 2.6.1 Exploring the simulation speed vs. accuracy tradeoff . . 30
  - 2.6.2 Simulation Speed Comparison . . . 32

2.6.3 Impact of NoC design on system performance 33
2.7 Conclusions 34
Bibliography 35

3 A Performance Analysis of Chiplet-Based Systems 39
3.1 Introduction 41
3.2 Overview of Chiplet-Based Architectures 42
3.2.1 Chiplet-based NoCs 43
3.2.2 NUMA-aware Memory Allocation 44
3.2.3 Last Level Cache Organization 44
3.2.4 Yield and Cost 45
3.3 Experimental Methodology 46
3.3.1 System Configuration 46
3.3.2 Simulation Setup 47
3.3.3 Workloads 48
3.3.4 Evaluated Systems 49
3.4 Performance Evaluation 49
3.5 Conclusions 53
Bibliography 54

Appendix A: Additional Results on Cost Analysis 57

4 PRISM: Reducing Performance Overheads of Multi-chiplet Interconnects 61
4.1 Introduction 63
4.2 Related Work 64
4.3 Design 66
4.4 Experimental Methodology 73
4.5 Evaluation Results 75
4.6 Conclusions 80
Bibliography 81

Chapter 1

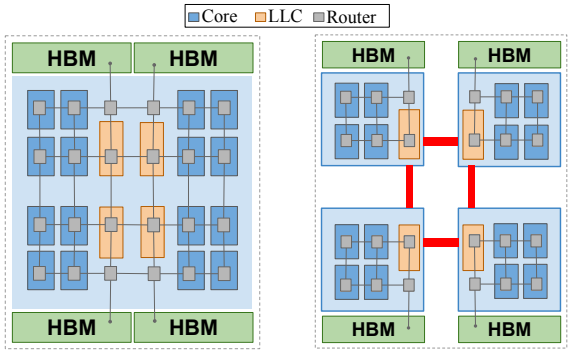
Introduction

Since the early 2000s, the semiconductor industry has relied on multicore scaling, increasing the number of cores and on-chip resources to continue delivering performance growth. At the same time, technology scaling has become increasingly challenging, and integrating more devices on a monolithic chip incurs substantial costs: large monolithic dies have low manufacturing yield, making them economically unviable at scale [1, 2]. Building chips out of multiple smaller chiplets offers a cheaper, higher-yield alternative while also enabling better scalability, modular design, and heterogeneous integration of components across different process technologies. In the past years, chiplet-based systems have been widely adopted by the industry, as seen in AMD's EPYC and RYZEN families [3]. More recent products, such as AMD MI300 [4] and Intel Sapphire Rapids [5], combine many CPU and GPU chiplets alongside high-bandwidth memory (HBM) nodes.

Despite their manufacturing advantages, multi-chiplet chips come with performance overheads. Due to their large size, such systems inevitably use multiple Non-Uniform Memory Access (NUMA) nodes, leading to performance bottlenecks that arise from varying memory access latencies. Beyond NUMA effects, inter-chiplet communication introduces further overheads: exchanging data between chiplets requires traversing longer links via a silicon interposer, adding extra latency, while the available off-chiplet bandwidth is inherently constrained by the limits of chiplet integration. Passive silicon interposers, often selected due to their lower cost [6], make these latency overheads more severe.

This thesis investigates the performance overheads of multi-chiplet systems, with a particular focus on their interconnection networks. It introduces a new, more scalable simulation infrastructure for studying large-scale multi-chiplet systems, measures and analyzes their performance overheads compared to monolithic designs, and proposes a network solution to mitigate inter-chiplet communication bottlenecks.

The remainder of the introductory chapter is structured as follows: Section 1.1 outlines the research problem, and Section 1.2 presents the thesis objectives and contributions.



(a) Monolithic System (b) Multi-chiplet System

Figure 1.1: Example of a 16-core monolithic chip and its equivalent four-chiplet design, showing cores, LLC banks, HBM controllers, and their on-chip network.

1.1 Problem Statement

Multi-chiplet chips have emerged as a practical response to the scaling limitations of monolithic designs: rather than a single large chip, the system is composed of multiple smaller chiplets interconnected through chiplet-to-chiplet links, as illustrated in Figure 1.1. This partitioning offers improved yield and lower manufacturing costs [1], but comes with performance trade-offs. Chiplet-based systems suffer from overheads related to remote memory accesses and inter-chiplet communication. This thesis focuses on the latter: data exchanged between chiplets must traverse longer interconnect paths with limited bandwidth, introducing latency and throughput constraints that directly limit system performance. These overheads are the core problem this thesis addresses. However, tackling them first required developing a simulation infrastructure capable of modeling chiplet systems at scale, then a systematic performance analysis to understand where performance is lost, and finally targeted network solutions to mitigate the identified bottlenecks.

1.1.1 The Simulation Scalability Challenge

Computer architecture research relies extensively on simulation, which enables the study and validation of new designs. Including Network-on-Chip (NoC) models into those simulations allows the assessment of how the on-chip interconnect affects overall system performance and exploration of the benefits of different NoC designs.

However, detailed microarchitectural simulation is computationally expensive, and the cost scales with system complexity. Simulators that prioritize accuracy, such as gem5 [7], become impractically slow for the scale at which chiplet systems operate. Faster alternatives, such as ZSim [8] and Sniper [9], forgo network modeling entirely to achieve their speed. Neither is well-suited for studying chiplet architectures, which require both simulation speed and interconnect accuracy. This thesis addresses this gap by developing a simulation framework that integrates accurate network modeling into a fast parallel simulator, enabling the large-scale chiplet studies that follow.

1.1.2 Performance Analysis of Chiplet-Based Systems

Despite the well-established cost and yield advantages of chiplet-based designs [1], their system performance overheads relative to monolithic chips had not been systematically studied. The design space is large: chiplet size, network bandwidth, last-level cache organization, and interposer type all influence system performance and interact in non-trivial ways. A comprehensive analysis covering these parameters is needed to measure and analyze the system performance overhead of a chiplet-based system. This thesis conducts such an analysis, revealing the inter-chiplet network as a primary performance overhead and establishing the motivation for the network solution presented next.

1.1.3 Reducing Inter-Chiplet Communication Overhead

Once the performance bottlenecks of chiplet-based systems are understood, addressing them requires targeted network solutions. The inter-chiplet network faces constraints that differ fundamentally from those of intra-chiplet networks: chip-to-chip links carry higher latency, bandwidth across chiplet boundaries is limited, and passive silicon interposers provide no active routing support [6]. Standard network topologies and routing algorithms, designed for monolithic chips, do not account for these constraints. Most proposals for multi-chiplet systems assume active interposer components that are absent in the passive interposer case [2, 10, 11]. A network solution that jointly addresses topology, routing, and bandwidth utilization under these constraints is therefore needed. This thesis proposes a new network design for multi-chiplet chips on passive silicon interposers, targeting the latency and bandwidth constraints of inter-chiplet communication.

1.2 Thesis Objectives and Contributions

1.2.1 Fast, Large-Scale Simulation with Detailed Interconnect Modeling

Objective:

The first objective of this thesis is to develop a microarchitectural simulator that combines the speed of fast parallel simulators with the detailed interconnect modeling of accurate network simulators.

Related Work:

Plenty of microarchitectural and system simulators exist. gem5 [7], one of the most widely used, offers detailed interconnect modeling, but it is sequential, which makes it relatively slow. ZSim [8] and Sniper [9] are parallel and orders of magnitude faster, but neither models on-chip interconnects. On the network simulation side, BookSim [12] is a widely used cycle-accurate simulator that is flexible in terms of topologies, allocators, and routing algorithms. Garnet [13], integrated into gem5, uses a more abstract pipeline approach, making it faster but less detailed than BookSim. Integrating either into a microarchitectural simulator substantially decreases simulation speed [14].

Analytical methods offer a lighter-weight alternative [15, 16], though their accuracy degrades significantly under network congestion.

Accurately modeling the network is increasingly important as NoC designs have evolved substantially, with recent high-performance designs such as ShortPath [17] and FastTrackNoC [18] approaching the theoretical minimum latency through fine-grained router bypassing. These designs are used in this work to assess the impact of high-performance NoCs on system performance.

Thesis Approach:

BookSim is integrated into ZSim to produce BZSim, a fast, large-scale microarchitectural simulator with detailed interconnect modeling. To make the network simulation lightweight, BZSim is based on the conjecture that the latency of low-contention traffic can be calculated analytically relatively accurately. Traffic is classified by its estimated level of contention using a user-defined threshold. For traffic below that threshold, latency is calculated analytically rather than simulated. This reduces the amount of traffic that needs to be simulated, lowering the workload of network modeling. In addition, individual parts of the network are monitored separately and their evaluation is skipped when they are not used. The combination of these two mechanisms offers a knob for trading simulation accuracy for speed.

Thesis Contributions: In line with this objective, **Paper I** developed BZSim and made the following contributions:

- Integrated BookSim, an accurate network simulator into ZSim, a fast parallel system simulator to produce BZSim. BZSim is an open-source fast, large-scale microarchitectural simulator with detailed on-chip interconnect modeling.
  - BZSim is an order of magnitude faster than an optimistic gem5 setup with point-to-point interconnects, and 3× slower than ZSim, which does not model on-chip networks.
- Developed a novel mechanism that identifies low-contention network traffic and skips its simulation using an analytically calculated latency. This enables a tunable tradeoff between simulation speed and accuracy.
  - BZSim skips network simulation for 20–85% of traffic depending on the configured threshold
  - BZSim achieves 2–4× speedup with a maximum error of 3–5% in normalized IPC and 10–20% in average packet latency.
- Modeled and evaluated FastTrackNoC, and ShortPath, two state-of-the-art high-speed NoC designs and their impact on system performance on a 32-core system.
  - FastTrackNoC on average improves system performance by 18% over a baseline NoC, reducing the average packet latency of the network to 49% of the baseline.
  - ShortPath on average improves system performance by 10% over a baseline NoC, reducing the average packet latency of the network to 69% of the baseline.

1.2.2 Performance Analysis of Chiplet-Based Systems

Objective:
The second objective of this thesis is to analyze the performance of chiplet-based systems and compare it to monolithic chips. The analysis covers a range of design parameters, including system and chiplet size, memory hierarchy organization, interconnect configuration, and technology aspects.

Related Work:
Commercial chiplet designs such as the AMD MI300 [4], and Intel Sapphire Rapids [5] demonstrate the industry shift away from monolithic dies. The cost and yield benefits of this approach have been thoroughly quantified [1].

Prior work has studied design choices that affect inter-chiplet networks, including the choice of silicon interposer [6,19], and inter-chiplet link placement and width [20–23]. The effect of memory organization on inter-chiplet traffic is equally significant: multi-chiplet chips are inherently NUMA [24–26], and both data placement and cache organization determine how much traffic must cross chiplet boundaries. Yet the combined effect of these design choices on the performance of a multi-chiplet system compared to a monolithic had not been systematically evaluated.

Thesis Approach:
Performance is evaluated by comparing three system configurations: a chiplet-based design, its monolithic counterpart of equivalent core count, and an ideal chiplet baseline that eliminates remote memory accesses to serve as a performance upper bound. The analysis covers several design parameters: system and chiplet size, LLC organization, intra-chiplet NoC datapath width, and silicon interposer type. For LLC organization, two configurations are studied: a sliced design, where the cache is distributed across chiplets, and a private design, where each chiplet maintains its own local LLC. BZSim is extended to support the modeling of chiplet-based architectures with detailed cycle-accurate interconnect and memory system models. The performance analysis is complemented by a cost evaluation using the Feng-Ma chiplet actuary model [1], to put the performance overheads in perspective against the cost benefits of chiplet-based designs.

Thesis Contributions: In line with this objective, **Paper II** conducted a systematic performance analysis and made the following contributions:

- Extended BZSim to model large-scale chiplet-based architectures with detailed cycle-accurate models of their interconnection network and memory system.
- Conducted the first thorough analysis of performance overheads and cost-performance tradeoffs of chiplet-based chips compared to monolithic. Despite costing only 55% as much, chiplet-based systems:
  - Achieve only 43–75% of monolithic IPC (58% on average), while an ideal chiplet system with perfect data locality outperforms monolithic by 24%.

  - Suffer 3.7× higher average packet latency and 40% higher average memory access time.

- Analyzed technological aspects that determine key system parameters, including inter-chiplet link lengths and delays, microbump budgets, and the yield and cost of chiplet-based versus monolithic chips, confirming that chiplet-based designs reduce recurring engineering costs by nearly half.
- Investigated the impact of key design parameters on system performance, including system size, chiplet size, LLC organization, NoC datapath width, and silicon interposer type.
  - Private LLC improves the system performance by about 30% compared to sliced LLC in multi-chiplet systems.
  - Active interposers recover most of the performance overhead, bringing multi-chiplet systems to within 3% of monolithic performance, but at a 61% higher cost due to their lower yield.
  - Wider intra-chiplet NoC links improve the network throughput, reducing average memory access time across both chiplet and monolithic configurations.
  - The performance gap remains relatively stable as system size increases.
  - For a constant system size, performance decreases with finer chiplet granularity: systems with 2, 4, and 8 cores per chiplet achieve 76%, 58%, and 44% of monolithic IPC respectively.

Appendix A of **Paper II** provides additional results on the cost analysis of various design options for chiplet-based systems, which were not included in the paper due to space constraints.

1.2.3 Reducing Inter-Chiplet Communication Overhead

Objective:

Paper II showed that inter-chiplet links are a key performance bottleneck in multi-chiplet systems. Crossing the chiplet boundaries has bandwidth limitations and latency overheads. The bandwidth limitation stems from the number of microbumps a chiplet can accommodate, which is constrained by the chiplet size and density allowed by the packaging technology. Moreover, cross-chiplet connections via a silicon interposer, especially a passive one, add extra delay. The third objective of this thesis is to alleviate these overheads by introducing a novel network design for multi-chiplet chips on passive silicon interposers.

Related Work:

Several works have explored network topologies for multi-chiplet systems based on active interposers [2, 10, 11, 20], which can embed routing logic but are significantly more expensive than passive alternatives [6]. For passive interposers, Feng et al. [27] showed that grouping inter-chiplet links at chiplet edges enables high-radix topologies, delivering up to 2× saturation throughput

and 45% lower latency than a conventional 2D-mesh multi-chiplet network. Network-on-Die (NoD) [28] moves all inter-chiplet routers onto a standalone routing chiplet, though this does not scale well as chiplet count increases.

Regardless of topology, combining intra- and inter-chiplet networks introduces deadlock risks. Common deadlock avoidance strategies include turn restriction [29, 30] and injection gating [31] which break dependency cycles but either reduce path diversity or throttle traffic under congestion. Alternatively, virtual-network (VN) separation offers better performance [32–34], but existing VN-based schemes target active interposers.

Compression offers a complementary direction to address the limited C2C bandwidth. Several designs compress all network traffic at every network interface or router [35, 36], adding latency overhead even at low network load. Selective approaches compress only congested paths to avoid this penalty [37–39]. All prior compression work, however, targets monolithic chips; no prior work has applied compression to inter-chiplet links.

Thesis Approach:

PRISM introduces a new network for multi-chiplet chips on a passive silicon interposer which alleviates the bandwidth limitations and latency overheads of inter-chiplet communication. It describes a new topology that uses dedicated edge routers at chiplet boundaries interconnected within each chiplet with bypass links to reduce hop count for inter-chiplet traffic. In addition, it utilizes more effectively the chiplet-to-chiplet (C2C) bandwidth by compressing inter-chiplet traffic. The compression decision is made dynamically per flit and chosen only when C2C congestion is detected, thereby minimizing compression overheads. A deterministic routing algorithm with virtual-network separation provides deadlock-free routing.

Thesis Contributions: In line with this objective, **Paper III** introduced PRISM and made the following contributions:

- A new topology for multi-chiplet chips with passive silicon interposer is introduced.
  - Dedicated edge routers at chiplet boundaries handle inter-chiplet traffic, avoiding interference with intra-chiplet traffic.
  - Peripheral bypass links connect edge routers within a chiplet, reducing hop count for inter-chiplet traffic.
- A deterministic deadlock-free routing algorithm for the above topology is developed by adapting a two-virtual-network separation scheme [33], originally designed for active interposers, to fit the constraints of passive silicon interposers.
- PRISM is the first to apply compression exclusively for C2C communication on multi-chiplet chips. Its edge routers are designed to selectively compress inter-chiplet traffic based on their latency criticality and router congestion, thereby increasing effective C2C bandwidth with reduced latency overheads.
  - 11–33% of network traffic crosses chiplet boundaries and is eligible for compression, achieving a compression ratio of 1.1–4.9×.

  - Selective compression skips compression in 18–44% of cases, improving system performance over always-compressing by up to 6.5%.
- PRISM is evaluated on a 256-core system with 16 chiplets, demonstrating:
  - Up to 1.9× higher peak network throughput.
  - Up to 5.8× reduction in inter-chiplet packet latency.
  - Up to 17.4% improvement in system performance (10.0% on average).

1.3 Thesis Outline

The rest of this thesis consists of the three included papers and is organized as follows. Chapter 2 contains **Paper I**, "*BZSim: Fast, Large-scale Microarchitectural Simulation with Detailed Interconnect Modeling*". Chapter 3 contains **Paper II**, "*A Performance Analysis of Chiplet-Based Systems*". Chapter 4 contains **Paper III**, "*PRISM: Reducing Performance Overhead of Multi-chiplet Chip Interconnects*".

Bibliography

[1] Y. Feng and K. Ma, "Chiplet Actuary: A Quantitative Cost Model and Multi-Chiplet Architecture Exploration," in *Proceedings of the 59th ACM/IEEE Design Automation Conference (DAC)*, Jul 2022, pp. 121–126.

[2] A. Kannan, N. E. Jerger, and G. H. Loh, "Enabling Interposer-based Disintegration of Multi-core Processors," in *2015 48th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, Dec 2015, pp. 546–558.

[3] S. Naffziger, N. Beck, T. Burd, K. Lepak, G. H. Loh, M. Subramony, and S. White, "Pioneering Chiplet Technology and Design for the AMD EPYC™ and Ryzen™ Processor Families: Industrial Product," in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*, Jun 2021, pp. 57–70.

[4] Advanced Micro Devices, Inc., "AMD CDNA™ 3 Architecture," 2023. [Online]. Available: https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/white-papers/amd-cdna-3-white-paper.pdf

[5] N. Nassif, A. O. Munch, C. L. Molnar, G. Pasdast, S. V. Lyer, Z. Yang *et al.*, "Sapphire Rapids: The Next-Generation Intel Xeon Scalable Processor," in *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, Feb 2022, pp. 44–46.

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

[7] N. Binkert, B. Beckmann, G. Black, S. K. Reinhardt, A. Saidi, A. Basu, J. Hestness, D. R. Hower, T. Krishna, S. Sardashti, R. Sen, K. Sewell, M. Shoaib, N. Vaish, M. D. Hill, and D. A. Wood, "The gem5 simulator," *SIGARCH Comput. Archit. News*, vol. 39, no. 2, pp. 1–7, 2011.

[8] D. Sanchez and C. Kozyrakis, "ZSim: Fast and accurate microarchitectural simulation of thousand-core systems," in *Proceedings of the 40th Annual International Symposium on Computer Architecture (ISCA)*, 2013, pp. 475–486.

[9] T. E. Carlson, W. Heirman, and L. Eeckhout, "Sniper: Exploring the level of abstraction for scalable and accurate parallel multi-core simulation,"

in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, 2011.

[10] H. Shabani and X. Guo, "Cluscross: A new topology for silicon interposer-based network-on-chip," in *Proceedings of the 13th IEEE/ACM International Symposium on Networks-on-Chip (NOCS)*, 2019.

[11] S. Bharadwaj, J. Yin, B. Beckmann, and T. Krishna, "Kite: A family of heterogeneous interposer topologies enabled via accurate interconnect modeling," in *2020 57th ACM/IEEE Design Automation Conference (DAC)*, 2020, pp. 1–6.

[12] N. Jiang, D. U. Becker, G. Michelogiannakis, J. Balfour, B. Towles, D. E. Shaw, J. Kim, and W. J. Dally, "A detailed and flexible cycle-accurate network-on-chip simulator," in *IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 2013, pp. 86–96.

[13] N. Agarwal, T. Krishna, L.-S. Peh, and N. K. Jha, "GARNET: A detailed on-chip network model inside a full-system simulator," in *IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 2009, pp. 33–42.

[14] I. Pérez, E. Vallejo, M. Moretó, and R. Beivide, "BST: A BookSim-based toolset to simulate NoCs with single- and multi-hop bypass," in *IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 2020, pp. 47–57.

[15] A. E. Kiasari, Z. Lu, and A. Jantsch, "An analytical latency model for networks-on-chip," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 21, no. 1, pp. 113–123, 2013.

[16] S. Y. Narayana, S. K. Mandal, R. Ayoub, M. M. Islam, M. Kishinevsky, and U. Y. Ogras, "Fast analysis using finite queuing model for multilayer NoCs," *IEEE Design & Test*, vol. 40, no. 6, pp. 112–124, 2023.

[17] A. Psarras, I. Seitanidis, C. Nicopoulos, and G. Dimitrakopoulos, "ShortPath: A network-on-chip router with fine-grained pipeline bypassing," *IEEE Transactions on Computers*, vol. 65, no. 10, pp. 3136–3147, 2016.

[18] A. Ejaz and I. Sourdis, "FastTrackNoC: A NoC with FastTrack router datapaths," in *IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2022, pp. 971–985.

[19] D. Stow, I. Akgun, and Y. Xie, "Investigation of Cost-Optimal Network-on-Chip for Passive and Active Interposer Systems," in *2019 ACM/IEEE International Workshop on System Level Interconnect Prediction (SLIP)*, Jun 2019, pp. 1–8.

[20] N. E. Jerger, A. Kannan, Z. Li, and G. H. Loh, "NoC Architectures for Silicon Interposer Systems: Why pay for more wires when you can get them (from your interposer) for free?" in *2014 47th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, Dec 2014, pp. 458–470.

[21] 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 *2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, Nov 2018, pp. 1–8.

[22] Y. Feng, D. Xiang, and K. Ma, "A Scalable Methodology for Designing Efficient Interconnection Network of Chiplets," in *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, Feb 2023, pp. 1059–1071.

[23] ——, "Heterogeneous Die-to-Die Interfaces: Enabling More Flexible Chiplet Interconnection Systems," in *Proceedings of the 56th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, Oct 2023, pp. 930–943.

[24] D. Lenoski, J. Laudon, K. Gharachorloo, W.-D. Weber, A. Gupta, J. Hennessy, M. Horowitz, and M. S. Lam, "The Stanford Dash Multiprocessor," *Computer*, vol. 25, no. 3, pp. 63–79, Apr 1992.

[25] J. Laudon and D. Lenoski, "The SGI Origin: A ccNUMA Highly Scalable Server," in *24th Annual International Symposium on Computer Architecture (ISCA)*, Jun 1997, pp. 241–251.

[26] F. Dahlgren and J. Torrellas, "Cache-Only Memory Architectures," *Computer*, vol. 32, no. 6, pp. 72–79, Jun 1999.

[27] Y. Feng, D. Xiang, and K. Ma, "A scalable methodology for designing efficient interconnection network of chiplets," in *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2023, pp. 1059–1071.

[28] M. Ebrahimi, A. Y. Weldezion, and M. Daneshtalab, "NoD: Network-on-die as a standalone NoC for heterogeneous many-core systems in 2.5D ICs," in *2017 19th International Symposium on Computer Architecture and Digital Systems (CADS)*, 2017, pp. 1–6.

[29] J. Yin, Z. Lin, O. Kayiran, M. Poremba, M. S. B. Altaf, N. Enright Jerger, and G. H. Loh, "Modular routing design for chiplet-based systems," in *Proceedings of the 45th Annual International Symposium on Computer Architecture (ISCA)*, 2018, pp. 726–738.

[30] W. Cao, S. Jiang, and L. Huang, "Fast turn restriction algorithm to build deadlock-free modular chiplet integration systems," in *2022 IEEE 4th International Conference on Circuits and Systems (ICCS)*, 2022, pp. 27–32.

[31] P. Majumder, S. Kim, J. Huang, K. H. Yum, and E. J. Kim, "Remote control: A simple deadlock avoidance scheme for modular systems-on-chip," *IEEE Transactions on Computers*, vol. 70, no. 11, pp. 1928–1941, 2021.

[32] E. Taheri, S. Pasricha, and M. Nikdast, "DeFT: A deadlock-free and fault-tolerant routing algorithm for 2.5D chiplet networks," in *Design, Automation and Test in Europe Conference (DATE)*, 2022, pp. 1047–1052.

[33] ——, "ReD: A reliable and deadlock-free routing for 2.5-D chiplet-based interposer networks," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 43, no. 12, pp. 4599–4612, 2024.

[34] D. Yu, A. Li, N. Jing, J. Jiang, W. Sheng, and Q. Wang, "VDA: A simple but efficient virtual-channel-based deadlock avoidance scheme for scalable chiplet networks," in *Proceedings of the Great Lakes Symposium on VLSI 2024*, 2024, pp. 357–363.

[35] J. Zhan, M. Poremba, Y. Xu, and Y. Xie, "Noδ: Leveraging delta compression for end-to-end memory access in NoC based multicores," in *2014 19th Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2014, pp. 586–591.

[36] D. Deb, R. M.K., and J. Jose, "Flitzip: Effective packet compression for NoC in multiprocessor system-on-chip," *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 1, pp. 117–128, 2022.

[37] Y. Jin, K. H. Yum, and E. J. Kim, "Adaptive data compression for high-performance low-power on-chip networks," in *2008 41st IEEE/ACM International Symposium on Microarchitecture*, 2008, pp. 354–363.

[38] F. Jiang, C. Li, L. Chen, J. Liu, W. Zhang, and J. Xu, "SCNoCs: An adaptive heterogeneous multi-NoC with selective compression and power gating," in *2024 29th Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2024.

[39] Y. Wang, H. Li, Y. Han, and X. Li, "A low overhead in-network data compressor for the memory hierarchy of chip multiprocessors," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 6, pp. 1265–1277, 2018.