Hierarchical Network-on-Chip Design for Interposer-Based Systems and DNN Accelerators

by

Hesam Eddin Shabani

A Dissertation

Presented to the Graduate and Research Committee

of Lehigh University

in Candidacy for the Degree of

Doctor of Philosophy

in

Electrical Engineering

Lehigh University

(May 2023)

Approved and recommended for acceptance as a dissertation in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Hesam Eddin Shabani

Hierarchical Network-on-Chip Design for Interposer-Based Systems and DNN Accelerators

_____

Defense Date

_____

Dissertation Director

_____

Approved Date

Committee Members:

_____

Prof. Wujie Wen, Committee Chair

_____

Prof. Xiaochen Guo

_____

Prof. Zhiyuan Yan

_____

Prof. Mahdi Nikdast

Acknowledgements

I would like to express my sincere gratitude to Prof. Guo, who allowed me to join the ECE-CompArch Lab research team and conduct my research under her supervision. She has kindly supported me during this project, and her insightful comments helped me to proceed with my research in the right direction leading to publications in prestigious conferences. I would also like to thank my committee members Prof. Wujie Wen, Prof. Zhiyuan Yan, and Prof. Mahdi Nikdast, for their suggestions and feedbacks.

"Dedicated to my beloved parents and my lovely brother" for their love, endless support, encouragement, and sacrifices.

Contents

List of Tables

List of Figures

Abstract

Network-on-Chip (NoC) is a crucial chip multiprocessor component to communicate between many nodes. Continued increases toward multicore and manycore scalability have led to performance challenges of NoCs because of the increasing network diameter. Also, up to 30% of the chip's overall power budget is contributed by NoCs in modern chips [25], and on-chip power consumption exceeds the total power budget by increasing cores in the general-purpose chip multiprocessors. The hierarchical design approach is a promising solution to offer straightforward paths to improve performance and minimize power consumption. Hierarchical on-chip interconnection design is suitable for large systems by providing routes with shorter hop counts in the network. The hierarchical design approaches generally require inter-chip communication; however, as the number of small chips increases, the chip-to-chip communication becomes a performance bottleneck. Therefore, the interconnection network should be carefully designed to provide the shortest paths for as many source-destination pairs and avoid network congestion and minimum area and power consumption overhead. Furthermore, many widespread applications like modern AI systems require a large amount of data to support the computation, creating considerable data movement for on-chip and off-chip communications. Therefore, general-purpose on-chip network designs could not be appropriate for providing power efficiency in large-scale AI systems. Application-specific on-chip networks are proposed to leverage in the embedded systems to address the mentioned challenges in the general purpose. Therefore, hierarchical interconnect approaches such as tile-based architectures are well

studied and applied frequently in deep-learning accelerator designs. In this dissertation, three projects have been proposed. The first work proposes a new hierarchical topology design, ClusCross, to improve multicore interconnection networks on silicon interposer-based systems. The key idea is to treat each small chip as a cluster and use cross-cluster long links to increase bisection width and decrease average hop count without increasing the number of ports in the routers. The second work proposes a HW/SW co-design architecture to compute SpGEMM efficiently without requiring complex interconnection networks. A novel fast-packing algorithm, SorPack, is proposed to convert a sparse matrix into a dense matrix that increases PE utilization. Additionally, The HIRAC, a novel hierarchical accelerator, is proposed for executing Sparse GEMM and provides a scalable system that maximizes the parallelism of the PEs. The last chapter presents the heterogeneous design approach that can be used in applications requiring both sampled SpGEMM and Highly SpGEMMs and efficiently covering the higher sparsity ranges. The proposed heterogeneous design achieves 24% faster runtime estimation over HIRAC for a dynamic sparse attention matrix extracted from a state-of-the-art sparse attention model layer.

Chapter 1

Introduction

The increasing number of cores challenges the scalability of chip multiprocessors due to the requirements of high compute throughput systems. The hierarchical design approach is a promising solution to offer straightforward paths to improve scalability.

For example, multi-chip-modules packaging approaches such as silicon interposer-based systems [48] apply the idea of disintegration by partitioning a large chip into multiple smaller chips and using silicon interposer-based integration (2.5D) or organic substrates to connect these smaller chips. Since a small chip's verification, logic, and physical design are more convenient than a large chip, design costs are reduced compared to a large monolithic die. Also, this approach can improve overall yield because a smaller chip has fewer components, resulting in less likelihood of catching defects. In addition, modularity provides multiple smaller chips instead of a big chip; hence a tiny defective chip can be replaced at a lower cost when re-integrated through interposers. Moreover, as the number of cores grows, Network-on-chip (NoC)'s performance becomes limited because of the increasing network diameter. Hence, hierarchical design for on-chip interconnection provides routes with shorter hop counts from source to destination, which is more proper for large systems.

The hierarchical design approaches generally require inter-chip communication; however, as the number of small chips increases, the chip-to-chip communication

becomes a performance bottleneck. Therefore, the interconnection network should be carefully designed to provide the shortest paths for as many source-destination pairs, avoid network congestion, and minimize area and power consumption overhead.

Meanwhile, on-chip power consumption exceeds the total power budget by increasing cores in the general-purpose chip multiprocessors. The reason is the limitation of the power delivery network and thermal dissipation capability. Moreover, many widespread applications like modern AI systems require a large amount of data to support the computation, creating considerable data movement for on-chip and off-chip communications. As a result, it causes more energy consumption because data movement can consume more energy than computation. Therefore, general-purpose on-chip network designs could not be appropriate for providing power efficiency in large-scale AI systems.

Furthermore, application-specific on-chip networks are proposed to leverage in the embedded domain, which generates on-chip interconnection corresponding to the application's communication graph to address the challenges in general-purpose one. For example, the interconnection network in the specialized accelerator's design, tailored to the dataflow in Deep Neural Networks (DNNs) applications, is essential to meet the system requirements. The goal is to enable a large-scale system and extract the maximum possible parallelism from the available processing elements (PEs). Therefore, hierarchical interconnect approaches such as tile-based architectures are well studied and applied frequently in deep-learning accelerator designs. [7], [8], [45], [48].

For instance, Simba [48] uses a hierarchical interconnection design consisting of a Mesh NoC topology and a network-on-package (NoP). The Mesh NoC connects multiple processing elements (PEs) efficiently on the same chiplet. The NoP connects chiplets on the same package to provide the design for a large-scale system.

Additionally, SIGMA [45] leverages hierarchical design to interconnect different PEs to make the maximum possible parallelism of the in-use multipliers. SIGMA design includes NoC design and a combination of Flex-DPE units to construct a

Flex-DPU, which each Flex-DPU is for running one general matrix-matrix multiplication (GEMM). Hence, Multiple Flex-DPUs can consider in parallel to run multiple GEMMs. Also, the NoC is responsible for providing interconnection among the Flex-DPEs like the idea of the other tile-based architectures [17], [48]. However, the SIGMA has a 37.7% area overhead and 82% more power consumption as compared to the TPU because of the high flexibility and complexity of interconnection networks in the non-blocking distribution and reduction networks. Additionally, the utilization of SIGMA is determined by the sparsity of the streaming matrix. Thereby, the SIGMA design is not performance efficient for a sparser streaming matrix.

Chapter 2

ClusCross

As the number of small chips increases in the hierarchal design of the interposer-based system, chip-to-chip communication becomes a performance bottleneck. Hence, the interconnection network design should be a target to improve system performance carefully.

This work proposes a new network topology, ClusCross, to improve multicore interconnection networks on silicon interposer-based systems. The key idea is to treat each small chip as a cluster and use cross-cluster long links to increase bisection width and decrease average hop count without increasing the number of ports in the routers. Synthetic traffic patterns and real applications are simulated on a cycle-accurate simulator. Network latency reduction and saturation throughput improvement report compared to previously proposed topologies. Two versions of the ClusCross topology are presented. One version of ClusCross has a 10% average latency reduction for coherence traffic compared to the state-of-the-art network-on-interposer topology, the misaligned ButterDonut. The other version of ClusCross has a 7% and a 10% reduction in power consumption as compared to the FoldedTorus and the ButterDonut topologies, respectively.

2.1 Introduction

As the number of transistors increases, more processor cores can be integrated into a Chip Multi-Processor (CMP) to boost the computation throughput. With the invention of High Bandwidth Memories (HBMs) [42], memory bandwidth can be significantly improved by connecting multiple 3D-stacked DRAMs to processor chips through silicon interposers to satisfy the overall demands from the processor's cores. Each processor core, however, might need to access multiple memory locations and the increased number of cores also escalate coherence traffic among the cores. The on-chip networks are facing fundamental challenges to enable the scalability of the CMPs and to satisfy both the coherence and memory traffic demands. Meanwhile, with the increasing number of cores, on-chip power consumption is about to exceed the total power budget due to the limitation of the power delivery network and thermal dissipation capability. On-chip network designs have to be power efficient to meet the system power constraint.

Inspired by silicon interposer-based memory integration (*e.g.*, HBM), which is also referred to as 2.5D integration, recent studies [32] proposed the idea of disintegration by taking apart a large system into smaller parts by using the interposer-based integration to improve overall yield. This is because a smaller chip has fewer components and is less likely to catch defects. Having multiple smaller chips instead of a big chip also provides modularity, and a small defective chip can be replaced at a lower cost when re-integrated through interposers. Nevertheless, as multiple smaller chips are integrated through the interposers, the amount of chip-to-chip communications increases. Heavy traffic through the interposers can become a performance bottleneck [32]. Moreover, any processor core can access different parts of the on-chip memories. Hence, the memory traffic also needs to pass across different chips through the interposers. Even though disintegration can improve the yield and reduce the fabrication cost, the interconnection network can become a performance bottleneck if it is not carefully designed to overcome the challenges posed by the interposer-based

multi-chip systems.

Topology is one of the most important elements in interconnection network design, which directly influences network performance. In interposer-based systems, memory traffic can compete with coherence traffic for bandwidth [32]. The network topology should be designed to reduce such contention by increasing the number of links and bandwidth on segments critical to both memory and coherence traffic.

This work proposes a new interconnection network topology, ClusCross, for silicon interposer-based multi-chip systems. This topology is based on the idea of clustering. In order to decrease the network diameter and increase the cross-chip bandwidth, ClusCross maps a cluster of routers onto each small chip and increases the number of cross-cluster long links. As a result, the proposed topology can increase path diversity and bisection bandwidth, which can help to reduce contentions between memory and coherence traffic. In addition, cross-cluster long links can effectively reduce the hop counts for long-distance communication in both memory and coherence traffic.

The main contributions of this work include the following:

- Two versions of ClusCross on-chip network topology are proposed to improve network performance in NoC-on-interposer systems through decreasing average hop count and increasing cross-chip bandwidth by leveraging long links.

- Performance and cost of the proposed ClusCross topologies are evaluated and compared against other existing topologies using synthetic memory and coherence traffic.

- System performance of ClusCross topologies is evaluated using the PARSEC suite traces appropriate for CMP assessment.

The rest of the chapter is organized as follows: In Section 2.2, a brief overview of the interconnection networks based on silicon interposers is provided, and related work for both conventional NoC topologies and topologies for silicon interposer systems is

discussed. Section 2.3 presents the structure of the ClusCross and two versions of this topology. In Section 2.4, evaluation results are shown using both synthetic traffic patterns and real applications. The proposed topologies are compared against other topologies designed for interposer-based systems. Section 2.5 concludes the work.

2.2 Background and Related Work

2.2.1 Interposer-Based Interconnection Networks

Technology scaling does not benefit wires as much as it does transistors [11]. On-chip communication becomes a bottleneck for both power consumption and performance. Three-dimensional (3D) integration promises to bring processing elements and memory components physically close to each other to reduce wire distance and overcome the communication bottleneck. True 3D integration, however, requires through-silicon vias (TSVs), which are complicated to implement on processor dies and might introduce severe thermal issues and die yield reduction [32], [14]. As an alternative, individual chips can be connected to the silicon interposer layer through micro-bumps. Hence, memory and processor chips can be connected through a layer of silicon interposers on a substrate die to increase memory bandwidth.

Since interposer integration does not need TSVs in the silicon interposer layer, higher die yield and additional routing capabilities are provided for the system [44]. In addition, interposer-based systems have lower manufacturing and R&D costs as compared to the true 3D integration [44]. Although the physical design of the interposer integration also has technology-related challenges, such as thermal management and pin assignment [59], these challenges are solvable in the near term [44]. Consequently, interposer-based systems are the most promising near-term solution for die-stacking integration. Several commercial products of interposer-based ICs are already on the market [38], [39]. For example, the HBM uses TSVs to integrate stacks of DRAM dies and connects the DRAM stacks to the processor die using silicon interposers. Multi-