FROM THE EDITOR-IN-CHIEF

Beyond Wires: The Future of Interconnects

Hsien-Hsin S. Lee, *Intel Corporation, Hudson, MA, 01749, USA*

The first time I paid attention to interconnects in computing systems was during my graduate studies in the early 1990s at the University of Michigan, where I conducted research in parallel computing under the supervision of Professor Ed Davidson. In Ed's research group, most of us focused on performance analysis methodology (which Ed called "a functional approach") and parallelization techniques to understand and to accelerate large-scale simulation applications. The two projects I worked on included a car crash simulator sponsored by a nearby automobile company and an Indian Ocean circulation simulator, both were targeted to run on the Kendall Square Research machines and IBM SP2, a distributed memory machine using message-passing for shared data communication. Performance overheads due to node-to-node interprocessor communication, particularly in the IBM SP2, were always a top priority left to application programmers to manage, hide, and mitigate.

THIS PARADIGM SHIFT REVOLUTIONIZED PROCESSOR DESIGN FROM MONOLITHIC SINGLE-CORE ARCHITECTURES TO MODULAR BUILDING BLOCKS, USHERING INTO THE ERA OF MULTICORE ARCHITECTURES.

Then, in 2001, a seminal paper written by Ho, Mai, and Horowitz—"The Future of Wires"—was published in the *Proceedings of the IEEE*, which offered an outlook into wire delays as well as their profound implications on chip architecture design amid escalating wire scaling challenges. The insight of this paper prompted processor architects to reconsider the role and impact of wires, especially the long wires traversing across a large chip. This paradigm shift revolutionized processor design from monolithic single-core architectures to modular building blocks, ushering into the era of multicore architectures.

Fast-forward three decades, similar communication challenges continue to expand, from on-chip wires, to interconnects in larger systems integrating multiple (heterogeneous) chiplets using fancy, advanced packaging technology, and all the way to package-to-package communication and beyond. For example, to expedite on-chip communication, technology such as sub-NUMA clustering was introduced to reduce the latency of data movements across a large CPU die. In the modern cloud computing era, service-oriented architecture based on microservices, tensor parallelism exploitation, or model sharding for artificial intelligence/machine learning (AI/ML) training and inferences, and so on, rely on reliable, low-latency, and energy-efficient interconnects to achieve efficient computation and deliver accountable services. Furthermore, it is not just the interconnect latency that matters; interconnect bandwidth also plays a critical role in ensuring sufficient data to sustain high computational throughput. As shown by a recent study, "AI and Memory Wall" (to be published in our next Special Issue on Hot Chips), over the last 20 years, while dynamic random-access memory bandwidth and the FLOPS offered by hardware have been substantially improved by 100 times and 60,000 times, respectively, interconnect bandwidth has only increased by a modest factor of 30. This considerable mismatch could lead to imbalanced designs and thereby hinder the maximal performance achievable for modern applications that are bandwidth-limited.

In this Special Issue, we bring to you five selected works[A1,A2,A3,A4,A5] from the 2023 Hot Interconnects Symposium (HotI30), which is also celebrating its 30th anniversary. I would like to extend my gratitude to Dr. Scott Levy and Dr. Whit Schonbein, who served as the journal chairs of the symposium and also acted as our guest co-editors for reviewing and selecting invited papers to appear this issue from the 2023 Hot Interconnects program. The research scopes of these selected works include industry solutions BlueField data processing units (DPUs) to accelerate both lossy and lossless compression algorithms, techniques for

0272-1732 © 2024 IEEE
Digital Object Identifier 10.1109/MM.2024.3373336
Date of current version 9 April 2024.

APPENDIX: RELATED ARTICLES

A1. Y. Li, A. Kashyap, Y. Guo, and X. Lu, "Compression analysis for BlueField-2/-3 data processing units: Lossy and lossless perspectives," *IEEE Micro*, vol. 44, no. 2, pp. 8–19, Mar./Apr. 2024, doi: 10.1109/MM.2023.3343636.

A2. R. Oliveira and A. Gavrilovska, "Comprex: In-network compression for accelerating IoT analytics at scale," *IEEE Micro*, vol. 44, no. 2, pp. 20–30, Mar./Apr. 2024, doi: 10.1109/MM.2023.3343498.

A3. L. Dai, H. Qi, W. Chen, and X. Lu, "High-speed data communication with advanced networks in large language model training," *IEEE Micro*, vol. 44, no. 2, pp. 31–40, Mar./Apr. 2024, doi: 10.1109/MM.2024.3360081.

A4. D. Abts and J. Kim, "Enabling artificial intelligence supercomputers with domain-specific networks," *IEEE Micro*, vol. 44, no. 2, pp. 41–49, Mar./Apr. 2024, doi: 10.1109/MM.2023.3330079.

A5. D. Das Sharma and S. Choudhary, "Pipelined and partitionable forward error correction and cyclic redundancy check circuitry implementation for PCI Express 6.0 and Compute Express Link 3.0," *IEEE Micro*, vol. 44, no. 2, pp. 50–59, Mar./Apr. 2024, doi: 10.1109/MM.2023.3328832.

A6. S. Levy and W. Schonbein, "Special Issue on Hot Interconnects 30," *IEEE Micro*, vol. 44, no. 2, pp. 6–7, Mar./Apr. 2024, doi: 10.1109/MM.2024.3373338.

A7. H. E. Sumbul, J.-s. Seo, D. H. Morris, and E. Beigne, "A fully digital and row-pipelined compute-in-memory neural network accelerator with system-on-chip-level benchmarking for augmented/virtual reality applications," *IEEE Micro*, vol. 44, no. 2, pp. 61–70, Mar./Apr. 2024, doi: 10.1109/MM.2023.3338059.

A8. J. J. Yi, "Analysis of historical patenting behavior and patent characteristics of computer architecture companies—Part IX: Patent families," *IEEE Micro*, vol. 44, no. 2, pp. 72–77, Mar./Apr. 2024, doi: 10.1109/MM.2024.3373342.

A9. S. Greenstein, "Party like it's 1999?" *IEEE Micro*, vol. 44, no. 2, pp. 78–80, Mar./Apr. 2024, doi: 10.1109/MM.2024.3372349.

offloading compression to SmartNIC to reduce critical path latency, performance characterization of distributed large language models (LLMs) over different interconnects and communication protocols, high-throughput domain-specific interconnection network for modern AI workloads, and low-latency forward error correction and cyclic redundancy check in PCIe 6.0 and CXL 3.0. Please read the Guest Editorial message[A6] by Dr. Levy and Dr. Schonbein to get a preview of these articles.

Additionally, we feature an article from Meta Reality Labs' researchers demonstrating a cutting-edge fully digital compute-in-memory accelerator architecture crafted to enhance energy efficiency for augmented reality/virtual reality applications.[A7] This article is followed by a Micro Law department piece[A8] by Dr. Joshua Yi, part of a series exploring the patenting behavior and patent characteristics of computer architecture companies. In the article, Yi analyzes the characteristics of patent families, particularly the number of issued patents for each patent family. Finally, do not miss the compelling Micro Economics departmental article[A9] by Prof. Shane Greenstein of Harvard Business School titled "Party like it's 1999?" In it, Greenstein draws a striking parallel between the current industry-wide tsunami of feverish enthusiasm triggered by generative AI—or "gold rush" in his own terms—and the infamous dotcom bubble and the telecom bubble of the late 1990s, which led to financial fiascos across the globe. Could history repeat itself? This article leaves the question to the readers' discretion.

HSIEN-HSIN S. LEE is an Intel Fellow at Intel Corporation, Hudson, MA, 01749, USA. Contact him at lee.sean@gmail.com.