FROM THE EDITOR-IN-CHIEF

Toward Disaggregated and Heterogenous AI Systems

Hsien-Hsin S. Lee, *Intel Corporation, Boxborough, MA, 01719, USA*

This May, two major technology events unveiled progress and innovations that will continue to shape the future landscape of artificial intelligence (AI) systems. In Taipei, COMPUTEX 2025, a premier technological showcase that captivates both industry insiders and the general public, once again saw Nvidia and its supply chain take center stage. All eyes were on founder Jensen Huang, who was widely expected to deliver headline-grabbing announcements. Meanwhile, on the other side of the globe, the Red Hat Summit 2025 in Boston has placed a strong emphasis on AI and cloud computing, with a noteworthy focus on the strategy and evolution of "hybrid cloud." This growing momentum toward hybrid cloud for AI workloads signifies profound implications for how future computing infrastructure will be deployed and managed.

Several announcements from these two events point to the growing realization of heterogeneous AI infrastructure as a foundation of hybrid cloud. At COMPUTEX, Nvidia introduced its new "NVLink Fusion" ecosystem, a strategic expansion to open up its already-successful NVLink interconnect for their GPUs to third-party devices. This initiative allows non-Nvidia CPUs and accelerators to integrate more seamlessly into Nvidia's AI computing ecosystem. Although not a fully open standard, NVLink Fusion offers Intellectual Property components to plug in other devices to interoperate with Nvidia's Grace CPU (for third-party accelerators) or Nvidia GPUs (for non-Grace CPUs), creating a unified and more flexible platform toward heterogeneous AI compute. Shortly after, at Red Hat Summit 2025, Red Hat announced llm-d (large language model-disaggregated), an open source project aimed at building a distributed AI inference framework. Leveraging vLLM, llm-d enables efficient AI inference across disaggregated compute resources, allowing for seamless scaling and optimization of AI services.

Why is disaggregation significant? Transformer-based LLMs typically consist of two phases: the compute-intensive "prefill" phase, which generates the first token based on self-attention, and the memory bandwidth-limited "decode" phase, which generates subsequent tokens in an autoregressive loop. Disaggregation allows these two phases to run independently on distinct—potentially heterogeneous—devices, improving overall service efficiency and total cost of ownership (TCO) through an intelligent resource allocation policy and an effective job shop scheduler. Furthermore, it enhances energy efficiency per request, addressing growing concerns about the environmental footprint as AI inference becomes increasingly pervasive. However, this flexibility comes with tradeoffs, including additional data movement across devices, such as transferring the key-value (KV) cache, an essential storage structure used to skip redundant self-attention computations.

The entire story was then streamlined by a concurrent announcement from Nvidia regarding its revamped inference framework *Dynamo* (formerly known as *Triton Inference Server*). In this update, Nvidia signaled its strong support for the "llm-d" community initiative through their own tool—Dynamo. As part of these efforts, Dynamo will integrate Nvidia's *NIXL* (Nvidia Interconnect Xfer Library) to provide high-efficiency, point-to-point data movement application programming interfaces, which are specifically designed to handle intermittent data transfers (e.g., KV caches) across disaggregated devices and varying memory and storage tiers.

Combined with the earlier NVLink Fusion announcement, a clear vision begins to emerge in which Nvidia is building a comprehensive, vertically integrated, and somewhat standardized infrastructure to support large-scale distributed AI inference services from the bottom (hardware) to the top (software). This end-to-end service architecture is designed to accommodate both Nvidia and non-Nvidia devices, orchestrated

0272-1732 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies.
Digital Object Identifier 10.1109/MM.2025.3575180
Date of current version 27 June 2025.

APPENDIX: RELATED ARTICLES

[A1] R. Aitken and L. Yang, "Special Issue on Hot Chips 2024," *IEEE Micro*, vol. 45, no. 3, pp. 6–7, May/Jun. 2025, doi: 10.1109/MM.2025.3572594.

[A2] J. J. Yi, "A review of *Wisconsin Alumni Research Foundation v. Apple*—Part IV," *IEEE Micro*, vol. 45, no. 3, pp. 97–102, May/Jun. 2025, doi: 10.1109/MM.2025.3573578.

[A3] S. Greenstein, "The scramble after breakthrough," *IEEE Micro*, vol. 45, no. 3, pp. 108–110, May/Jun. 2025, doi: 10.1109/MM.2025.3567048.

[A4] G. Tyson, "Sally A. McKee," *IEEE Micro*, vol. 45, no. 3, p. 112, May/Jun. 2025, doi: 10.1109/MM.2025.3572559.

under Nvidia's Dynamo framework to deliver high performance at an optimized TCO. Furthermore, NVLink Fusion, if widely adopted, will solidify NVLink as the de facto standard for "scale-up" data communication in the future—potentially marginalizing effort from competing workgroups such as Ultra Accelerator Link (UALink) and Broadcom's recently proposed Scale-Up for Ethernet (SUE), both of which may face steeper challenges in gaining market traction.

There is a strong incentive for accelerator companies—particularly start-ups eagerly seeking their first foothold in the market—to join this fledging but resourceful ecosystem as it offers a pathway to greater hardware adoption while enabling peaceful coexistence with Nvidia. For those choosing to opt out, their journey is certainly more challenging: building an entirely independent infrastructure, both hardware and software, to compete end to end with Nvidia's formidable alliance. Overall, I find this strategy intriguing in achieving large-scale distributed AI inference, and its continued development will be important to watch.

THIS SPECIAL ISSUE HIGHLIGHTS SELECTED WORKS PRESENTED AT THE HOT CHIPS 2024 SYMPOSIUM.

This special issue highlights selected works presented at the Hot Chips 2024 symposium. First, I would like to express my sincere gratitude to our guest coeditors, who also served as the program co-chairs for Hot Chips 2024: Dr. Rob Aitken and Larry Yang, for making this special issue possible with impeccable quality. Hot Chips is a highly acclaimed flagship venue where the latest innovations and product previews in processors, accelerators, and their enabling technologies are unveiled to the community. It is our privilege to partner with them and publish selective presentations. In this issue, we feature nine notable products and implementations from Hot Chips, spanning different computing domains: x86 processors from Intel for client and data center computing, AI accelerators for edge and data centers from AMD, an ARM design from Qualcomm, an open source RISC-V implementation from the Chinese Academy of Sciences, a mainframe processor from IBM, an AI inference chip from FuriosaAI, and a tensor algebra accelerator from Stanford University. For a preview of these articles, we invite you to read the guest co-editor's introduction message.[A1]

In the fourth installment of *Wisconsin Alumni Research Foundation v. Apple* in the *Micro Law* column,[A2] Joshua Yi discusses the patent inventors' potential "inequitable conduct," a serious allegation in patent law, contested by Apple but ultimately dismissed by the judge. In the *Micro Economics* column,[A3] Prof. Shane Greenstein examines two strategic approaches to commercializing new technologies: incremental versus ambitious, using recent developments in generative AI services to exemplify.

This issue also offers a moment to honor the memory of Prof. Sally A. McKee—a dear friend in the community and my former collaborator. Together, we co-authored my first U.S. National Science Foundation-funded project, which was pivotal in launching both of our academic careers at Cornell and Georgia Tech in 2003. Prof. McKee's seminal paper "Hitting the Memory Wall: Implications of the Obvious" co-authored with her thesis advisor, the late National Academy of Engineering President Prof. Wm. A. Wolf, christened the term *memory wall*—a concept widely used today that looms more critical than ever in evaluating overall system performance in today's AI era. I would also like to thank Prof. Gary Tyson for writing a tribute to commemorate Prof. McKee.[A4]

I hope that you enjoy the topics selected for this issue.

HSIEN-HSIN S. LEE is an Intel Fellow at Intel Corporation, Boxborough, MA, 01719, USA. Conact him at lee.sean@gmail.com.