FROM THE EDITOR-IN-CHIEF

The Path to Powering Intelligence

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

Since the last major Internet revolution more than a quarter century ago, which brought the world search engines, e-commerce, and streaming services, we are once again marching into another transformative era, signified by the rapid expansion of computing infrastructures. A new global phenomenon is unfolding, driven largely by hyperscalers that build new data centers to power diverse artificial intelligence (AI) services across different industries. In recent news, Black Rock, Microsoft, and the United Arab Emirates's investment company MGX announced their partnership, dubbed the *Global AI Infrastructure Investment Partnership*, seeking to raise $30 billion from investors, and financing total capital up to $100 billion in data center and power infrastructure construction. Larry Ellison of Oracle disclosed an ongoing plan to expand its existing 66 data centers and build 100 more new data centers. Meanwhile, China's data center market is projected to reach $77 billion by 2029. Undoubtedly, AI is set to become pervasive and ubiquitous in this ongoing digital revolution, promising higher-quality services, enhanced user experiences, increased productivity, and a fundamental transformation of the way we live. However, this rapid development introduces significant new challenges when it comes to computing requirements, particularly the massive energy demands and their associated environmental impact, such as carbon emissions, which can be unsustainable.

To ensure nondisruptive services, hyperscalers are also exploring their own alternative solutions. For example, Oracle announced that it has secured permits to build three small modular reactors to meet the power needs of its AI data centers. It is likely that other large data center operators will follow suit as multigigawatt power requirements become essential for them to guarantee high-availability services. According to Electric Power Research Institute,[1] data centers in the United States are projected to consume up to 9.1% of the nation's electricity by 2030, a significant increase from 4% in 2023. Also, each AI query, such as prompts processed by ChatGPT's large language model, is estimated to use 10 times more electricity than a typical Google search query. The demand will only intensify when multimodal AI models that generate images and videos in response to text prompts become popular.

On the other hand, running AI services in a way that adheres to ethical standards and protects users' privacy will demand even more computational resources. Implementing the latest techniques to enable confidential AI end to end and AI guardrails on top of existing machine learning algorithms will increase computational loads, further straining energy consumption and exacerbating the environmental impact.

> SOVEREIGNTY AI, *A NEW FORCE, IS FERVENTLY SOUGHT AFTER BY GOVERNMENTS TO SAFEGUARD A NATION'S AUTONOMY, SECURITY, AND STRATEGIC INTERESTS.*

As a nation's power and leadership are increasingly defined by its overall compute capabilities, the resources for AI model development and deployment have become the key points of contention. Although such rivalries can accelerate technological developments in general, they also escalate tensions among countries, leading to increased, undesired confrontation. Possessing the most advanced computing infrastructure and AI techniques is no longer merely a matter of technology superiority, it now plays a critical role in national security and influences a country's ability to dominate and shape the global world order. *Sovereignty AI*, a new force, is fervently sought after by governments to safeguard a nation's autonomy, security, and strategic interests.

This special issue features seven invited technical articles that offer learnings, insights, perspectives, and highlight challenges of "The Past, Present, and Future of Warehouse-Scale Computing." First of all, I would like to thank our Associate Editor Prof. Gabriel Falcão for initiating the discussion about the possibility of putting together this special issue, partly as a tribute to Dr. Luiz André Barroso, the former fellow and VP of

0272-1732 © 2024 IEEE
Digital Object Identifier 10.1109/MM.2024.3467528
Date of current version 14 October 2024.

APPENDIX: RELATED ARTICLES

A1. J. L. Hennessy, C. Kozyrakis, and G. Falcão, "Special Issue on the Past, Present, and Future of Warehouse-Scale Computing," *IEEE Micro*, vol. 44, no. 5, pp. 6–7, Sep./Oct. 2024, doi: 10.1109/MM.2024.3467468.

A2. J. L. Hennessy, "Luiz André Barroso: Brilliant engineer, humble leader, and mentor," *IEEE Micro*, vol. 44, no. 5, pp. 8–10, Sep./Oct. 2024, doi: 10.1109/MM.2024.3456892.

A3. A. Bersatti, E. Kim, and H. Kim, "Quantifying CO_2 emission reduction through spatial partitioning in deep learning recommendation system workloads," *IEEE Micro*, vol. 44, no. 5, pp. 75–82, Sep./Oct. 2024, doi: 10.1109/MM.2024.3373443.

A4. J. J. Yi, "Analysis of historical patenting behavior and patent characteristics of computer architecture companies—Part XII: Patent families," *IEEE Micro*, vol. 44, no. 5, pp. 83–88, Sep./Oct. 2024, doi: 10.1109/MM.2024.3467488.

A5. S. Greenstein, "Commercial and scientific prototypes," *IEEE Micro*, vol. 44, no. 5, pp. 90–92, Sep./Oct. 2024, doi: 10.1109/MM.2024.3441788.

Engineering at Google, who passed away on 16 September 2023. Barroso was a pioneer in modern data center design. His significant contributions have profoundly defined and impacted today's warehouse-scale computing infrastructure. Once the concept for this special issue became concrete, Prof. Falcão teamed up with Prof. Christos Kozyrakis and Prof. John L. Hennessy of Stanford University to serve as guest co-editors for this special issue. I was impressed by how quickly they were able to enlist top researchers from leading companies and universities, including Alibaba, Amazon, Ecole Polytechnique Fédérale de Lausanne, Google, Meta, Microsoft, and the Massachusetts Institute of Technology, to contribute to this special issue. Also, this issue would not have been possible without the assistance of Dr. Parthasarathy (Partha) Ranganathan, another fellow and VP of Engineering at Google, who collaborated closely with Barroso. Barroso and Ranganathan previously served as the guest editors for an *IEEE Micro* Special Issue on Data Center-Scale Computing 14 years ago. Ranganathan provided valuable input from his experiences and suggested the title for the current special issue during the early planning phase. Please read the guest editorial message by Hennessy et al.[A1] for brief introductions of these seven articles. Finally, I would like to express my gratitude to Prof. Hennessy,[A2] chairperson of Alphabet Inc. and former president of Stanford University, who wrote a beautiful commemorative piece in memory of Barroso.

In addition to the invited articles, we feature a related article by Bersatti et al.[A3] from the Georgia Institute of Technology, which discusses a carbon footprint reduction strategy for deep learning recommendation systems. Personalized recommendation systems are widely employed in online services and consume a substantial amount of computing cycles for both training and inferences in commercial data centers such as those of Alibaba, Amazon, Google, Meta, Netflix, and so on. This article analyzes the impact of spatial partitioning of workloads that can decouple the geographical constraint and reduce the carbon footprint for recommendation models. Following this, Dr. Joshua Yi[A4] continues his sequel of analyzing the characteristics of patent families filed between 1996 and 2020 from 18 leading computer architecture companies for the regular Micro Law column. In the Micro Economics column, Prof. Shane Greenstein[A5] examines the distinction and intertwined relationship of commercial and scientific prototypes and how modern digitization revolutionized the prototyping process. He used ImageNet's and ChatGPT's development as illustrations in this article.

This Special Issue on the Past, Present, and Future of Warehouse-Scale Computing is timely, and I hope you enjoy the latest developments discussed in these articles. Its underlying infrastructure will continue to play a central role and become an integral part of our daily digital lives.

REFERENCE

1. "Powering intelligence: Analyzing artificial intelligence and data center energy consumption," Elect. Power Res. Inst., Palo Alto, CA, USA, May 2024. [Online]. Available: https://www.epri.com/research/products/000000003002028905

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