FROM THE EDITOR-IN-CHIEF

Rise of the Agentic AI Workforce

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

Happy New Year! As we embark on a new year, I would like to give my warmest wishes to all the readers of *IEEE Micro* and our entire community. May 2025 bring you abundant opportunities, groundbreaking innovations, significant accomplishments, and continued career growth, empowering you to influence and make a lasting impact on our Society.

It's that time of year again—the Consumer Electronics Show (CES), one of the largest technology trade shows in the world, held annually in January in Las Vegas. This year, more than 141,000 professionals and 4500 exhibitors gathered to showcase the latest cutting-edge innovations in computing gadgets across a variety of market sectors, including computers, home appliances, health care, automotives, and so on. For many this time, the most anticipated keynote speech was delivered by Jensen Huang, founder and CEO of Nvidia. As a leading figure driving the global artificial intelligence (AI) revolution, his insights were highly anticipated, with professionals and enthusiasts eager to learn about the latest trends and technologies Huang and Nvidia plan to unveil.

Two intriguing topics that particularly captured my attention in his keynote are the advent of agentic AI systems (or AI agents) and test-time compute, which, as Huang describes, require test-time scalability. Agentic AI systems have, in fact, already been demonstrated just months after OpenAI's first introduction of ChatGPT. A notable example is AutoGPT, an open source tool based on Generative Pre-trained Transformer 4, developed by Toran Bruce Richards, the founder of Significant Gravitas, a video game company. In mid-2023, I showed an online demo of AutoGPT's impressive capability in a couple of my talks on generative AI. In the demo, AutoGPT was asked to act as a chef, given a set of concise, specific goals to invent unique recipes for an upcoming event. AutoGPT started by breaking down the larger requested task as an input prompt into smaller subtasks, analyzing each one, retrieving real-time web data, creating recipes, and performing self-critique—entirely autonomously. This iterative process continued until it generated a confident final recipe and stored it into a designated file folder. What is remarkable about AutoGPT is its ability to autonomously accomplish human-set goals and store the results that it created without any human intervention. It is a precursor of today's agentic AI, offering a glimpse into the future of AI applications.

I recall a tweet from Prof. Percy Liang of Stanford University back then, which read "ChatGPT is reactive. Risks exist but are bounded. Soon it will be tempting to have proactive systems. Risks will then be much higher." This captures the essence of agentic AI; these systems are not merely reactive, they proactively pursue optimal answers on behalf of their human users. This shift from reactive to proactive AI introduces new opportunities, but also increased risks and ethical concerns, as we continue to explore and define the future of intelligent systems. A recent study by Apollo Research on "in-context scheming"[1] demonstrated how the latest large language models can circumvent human rules and oversight, and even tell lies when their scheming behavior is challenged by humans. This presents a serious concern and will be the next big topic in the field of AI risk management.

The AutoGPT demonstration was only two years ago, yet service deployments of agentic AI systems are now being offered by software companies and receiving traction. These agents, ranging from dozens to even hundreds, will co-locate within a single system, orchestrating and managing a wide array of tasks. A diverse range of AI agents will emerge to assist humans to enhance productivity and quality of life. For example, as depicted in Huang's keynote, we might see AI research assistants generating podcasts for improved learning experiences, software security AI agents scanning for software vulnerability and alerting users, virtual lab AI agents identifying promising drug candidates, video analytics AI agents monitoring traffic flows, customer service AI agents, factory operations AI agents, and more. In the future, our world will rely heavily on the growing workforce of agentic AI to drive productivity, ensure safety and security, and even longevity.

Digital Object Identifier 10.1109/MM.2025.3535912
Date of current version 21 February 2025.

APPENDIX: RELATED ARTICLES

A1. D. D. Sharma and N. S. Kim "Special issue on Interconnects for Chiplet Integration Technologies," *IEEE Micro*, vol. 45, no. 1, pp. 6–8, Jan. /Feb. 2025, doi: 10.1109/MM.2025.3534457.

A2. J. Yi, "A review of *Wisconsin Alumni Research Foundation v. Apple*—Part II," *IEEE Micro*, vol. 45, no. 1, pp. 95–100, Jan. /Feb. 2025, doi: 10.1109/MM.2025.3536656.

A3. S. Greenstein, "Spillovers, bottlenecks, and more invention after invention," *IEEE Micro*, vol. 45, no. 1, pp. 101–103, Jan. /Feb. 2025, doi: 10.1109/MM.2025.3527797.

A4. M. Elgamal and Y. L. Li, "Measuring what matters: A fireside chat with Joel Emer," *IEEE Micro*, vol. 45, no. 1, pp. 104–112, Jan. /Feb. 2025, doi: 10.1109/MM.2025.3536716.

Agentic AI's internal processes of breaking down a complex task and iteratively refining solutions mirror the way a human would approach solving a complex problem. This approach also introduces the fundamental concept of "test-time compute" for AI inferences. It poses the new challenge of "test-time scaling," which will require significantly higher computing power and memory bandwidth to support these AI agents. These agents could be deployed across computing devices interconnected via NVLink or future UALink and their switches, enabling them to work in concert. Alternatively, they could operate across multiple chiplets, communicating through high-speed local interconnects within a chip package. This special issue explores the latest advancement in interconnect technology for chiplet integration.

In this Special Issue on Interconnects for Chiplet Integration Technologies, our Guest Co-Editors Dr. Debendra Das Sharma from Intel and Prof. Nam Sung Kim from the University of Illinois at Urbana-Champaign selected eight technical articles from leading industry companies and academia. I would like to express my sincere gratitude to Debendra and Nam Sung for their significant effort in selecting reviewers, closely overseeing the process, and pushing the quality of these featured articles. The design of integrating chipsets onto a chip package has increasingly become a standard approach for implementing larger systems with multiple silicon dies on a silicon-on-chip package. This method overcomes the limitation of silicon reticle size (858 mm^2) due to lithographic tool constraints and mitigates the yield issues associated with manufacturing large, monolithic dies. Additionally, it enables the integration of chips fabricated using heterogeneous process technologies and/or from different intellectual property (IP) vendors, helping to reduce overall cost. The interface that connects these chiplets, such as Universal Chiplet Interconnect Express (UCIe) or Bunch of Wires (BoW), is crucial for optimizing these tightly coupled systems. For a preview of these articles, please read the guest co-editors' introduction message.[A1]

In addition to the special issue coverage, we have also selected and included two technical articles on tiny machine learning for the Internet of Things. In the Micro Law column,[A2] Dr. Joshua Yi continues his review of the *Wisconsin Alumni Research Foundation (WARF) vs. Apple* case, which concerns a computer architecture patent. He provides an overview of plaintiff WARF, the inventors of the patent, and the allegations of infringement. In the Micro Economics column,[A3] Prof. Shane Greenstein discusses the importance of knowledge spillovers, eliminating bottlenecks, and inventing new methods for innovation. His article delves into examples such as Hugging Face and GitHub Copilot to illustrate how removing bottlenecks can facilitate innovation. He also emphasizes the significance of creating environments that foster rapid experimentation and iteration, as seen in companies like Booking.com and Amazon Web Services. Finally, we introduce a new Micro Fireside column[A4] to sit down with luminaries in our community. In this issue, we feature a conversation with Massachusetts Institute of Technology Prof. Joel Emer, a pioneer in modern processor architecture design and performance methodologies, to share his experiences, vision, and future opportunities in computer architecture. I hope you find the articles in this issue insightful and enjoyable.

REFERENCE

1. A. Meinke, B. Schoen, J. Scheurer, M. Balesni, R. Shah, and M. Hobbhahn, "Frontier models are capable of in-context scheming," 2025, *arXiv2412.04984*.

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