FROM THE EDITOR-IN-CHIEF

A 33-GW Pilgrimage to the Promised Land of Agent Workforce

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

Imagine a near-future workplace where employers no longer need to hire humans for day-to-day business operations. Instead, artificial intelligence (AI) agents—an entirely new species in the modern (or virtual) office—carry out most of the tasks that real people were once paid to perform. These agents never tire or sleep, work around the clock without weekends or holidays, and operate with almost no mistakes. They do not negotiate for better benefits, organize unions to demand pay raises, or go on strike to disrupt productivity. What sounds like the opening chapter of a science fiction novel is now inching closer to reality. This new working norm, I believe, reflects the workplace of a not-so-distant future envisioned by many frontier thought leaders in AI.

Over the past couple of months, Sam Altman, CEO of OpenAI, has been making news by steadily forging partnerships, one after another, across the leading AI infrastructure supply chain, including vendors such as Oracle, Nvidia, AMD, Broadcom, and others. Thus far, OpenAI has announced future access to at least 33 GW of GPU capacity, an astonishing amount of computing power secured to support the AI services that OpenAI believes it will need in the coming years. Critics remain skeptical of these widely publicized deals. OpenAI is projected to generate roughly $20 billion in revenue this year with rosy forecasts that will reach hundreds of billions by 2030.[1] However, how, then, can the company afford to commit to deals whose cumulative value is well more than $1 trillion at today's prices?

Several of these agreements appear to involve circular revenue flows. Nvidia, for instance, has pledged $100 billion to fund AI infrastructure for OpenAI—capital that OpenAI is expected to spend on purchasing Nvidia's GPUs in return. As such, this circular capital flow artificially inflates Nvidia's revenue numbers, raising big questions of its legitimacy. AMD also announced a similar partnership shortly thereafter. Instead of a direct investment, AMD granted OpenAI stock warrants of up to 160 million shares, which will vest over time contingent on specific milestones to be met by OpenAI, including purchasing and deploying 6 GW of AMD GPU capacity in OpenAI's data center fleet in coming years. Yet, unlike the alleged circular economics of the Nvidia arrangement, this deal, I think, provides AMD with a strategic advantage: it gives AMD a great opportunity to have OpenAI, the most powerful AI private company today, validate and certify its GPU technologies. This is particularly meaningful for its ROCm (Radeon Open Compute platform) software framework, which has been facing an uphill battle against the formidable dominance of Nvidia's CUDA. Should OpenAI fulfill its terms, AMD will gain credibility via serving OpenAI's customers and become a viable alternative to those companies that are seeking a second GPU supplier to run AI on-premises or provide AI services. It could rapidly catalyze AMD's expansion of its GPU market share in the AI data center segment.

But, returning to the 33-GW GPU capacity; if Sam Altman is thinking about what I suspect he is, then such aggressive stockpiling of computing power may not seem completely unrealistic. A new wave of compute-intensive digital behavior is already emerging. For instance, Meta recently launched Meta Vibes through the Meta AI app, a platform where any user can create and directly share entirely AI-generated short videos. This new social network has gone viral, functioning as an AI-native counterpart to Instagram Reels and TikTok. Anyone with a mobile device can easily consume significant GPU cycles to generate media content using AI. More consequentially, just 10 months after my discussion on the rise of AI agents,[2] autonomous agentic AI systems have gained more traction in practical, commercial applications. In the coming era, companies are likely to hire armies of AI agents from providers like OpenAI to replace large portions of their human workforce. These AI agents will be hired and deployed on demand, work 24/7, and require none of the fringe costs for perks, health insurance, and retirement plans. In other words, the

Digital Object Identifier 10.1109/MM.2025.3637948
Date of current version 24 December 2025.

APPENDIX: RELATED ARTICLES

A1. W. Choi and J. Zhang, “Special Issue on Cache Coherent Interconnects and Resource Disaggregation Techniques,” *IEEE Micro*, vol. 45, no. 6, pp. 6–7, Nov./Dec. 2025, doi: 10.1109/MM.2025.3627696.

A2. J. J. Yi, “A review of *Wisconsin Alumni Research Foundation v. Apple*—Part VII,” *IEEE Micro*, vol. 45, no. 6, pp. 119–123, Nov./Dec. 2025, doi: 10.1109/MM.2025.3638484.

A3. S. Greenstein, “Private returns on technology adoption,” *IEEE Micro*, vol. 45, no. 6, pp. 124–126, Nov./Dec. 2025, doi: 10.1109/MM.2025.3615287.

agent service companies could effectively become the “boss” of most workers at companies across the globe. The computational demand required to power such AI workforces, token throughput, task processing, problem solving, continuous workloads, and more would be unimaginably enormous. In that light, even the 33 GW that Sam Altman has secured may one day seem small.

The initial discussion of the topic of this special issue began in early 2024 from my conversation with Prof. Myoungsoo Jung from Korea Advanced Institute of Science and Technology (KAIST). I knew Prof. Jung back in 2009 when he was working at Samsung Electronics while pursuing a graduate degree jointly bestowed by Georgia Tech and Korea University. He has been incredibly successful in multiple computer architecture areas including storage, nonvolatile memory, interconnect technologies, and processing-in-memory during his tenure as a faculty member at the University of Texas at Dallas, Yonsei University, and now an Endowed Chair Professor at KAIST. He is also a key contributing member to the Compute Express Link (CXL) and Ultra Accelerator Link Consortiums. Our conversation back then was centered around the CXL Intellectual Property (IP) and solutions under development by Panmnesia, a start-up company that he started and serves as the CEO. We envisioned that CXL would be a critical enabling standard to enable composable data center architecture and address the huge memory capacity demand by modern, fast-evolving AI applications. Therefore, we reached the decision to solicit recent works for a Special Issue on Cache-Coherent Interconnects and Resource Disaggregation Techniques and appointed Prof. Wonil Choi from Hanyang University and Prof. Jie Zhang from Peking University to serve as guest co-editors. I truly appreciate Prof. Choi and Prof. Zhang for their dedication and diligence to select the 12 outstanding articles featured in this issue. These works were contributed by authors from both academia and industry. For more details, please read the guest editors’ introductory message,[A1] which categorizes these works into three main areas: disaggregated storage, disaggregated memory, and interconnect.

As usual, in Part VII of the *Wisconsin Alumni Research Foundation (WARF) v. Apple* series for the Micro Law column,[A2] Dr. Joshua Yi focuses on two critical discovery disputes in the litigation. These disputes illustrate the broader challenges that courts face when balancing meaningful discovery with highly complex and rapidly evolving CPU products. In the Micro Economics column,[A3] Prof. Shane Greenstein reviews the lessons learned from previous successful adoptions of consumer computer technologies (CCTs), including mobile devices and the Internet and the related incremental and novel co-invention to utilize CCTs in business and create values. The article speculates that the same learning will be repeated in the new AI era.

I hope that you enjoy the final issue of 2025. Thank you again for your continued support of *IEEE Micro*. We wish you all a very Merry Christmas, and see you in 2026.

REFERENCES

1. A. Capoot, “Sam Altman says OpenAI will top $20 billion in annualized revenue this year, hundreds of billions by 2030,” *CNBC*, Nov. 6, 2025. [Online]. Available: https://www.cnbc.com/2025/11/06/sam-altman-says-openai-will-top-20-billion-annual-revenue-this-year.html
2. H.-H. S. Lee, “Rise of the agentic AI workforce,” *IEEE Micro*, vol. 45, no. 1, pp. 4–5, Jan./Feb. 2025, doi: 10.1109/MM.2025.3535912.

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