COMPUTING edge

- High Performance Computing
- Memory
- Careers
- Low Code

OCTOBER 2025

www.computer.org

IEEE COMPUTER SOCIETY

IEEE

COMPUTING edge

IEEE COMPUTER SOCIETY IEEE

IEEE COMPUTER SOCIETY computer.org

Printed with inks containing soy and/or vegetable oils

STAFF

Editor
Lucy Holden

Periodicals Portfolio Senior Managers
Carrie Clark and Kimberly Sperka

Director, Periodicals and Special Projects
Robin Baldwin

Production & Design Artist
Carmen Flores-Garvey

Periodicals Operations Project Specialists
Priscilla An and Christine Shaughnessy

Senior Advertising Coordinator
Debbie Sims

Circulation: *ComputingEdge* (ISSN 2469-7087) is published monthly by the IEEE Computer Society. IEEE Headquarters, Three Park Avenue, 17th Floor, New York, NY 10016-5997; IEEE Computer Society Publications Office, 10662 Los Vaqueros Circle, Los Alamitos, CA 90720; voice +1 714 821 8380; fax +1 714 821 4010; IEEE Computer Society Headquarters, 2001 L Street NW, Suite 700, Washington, DC 20036.

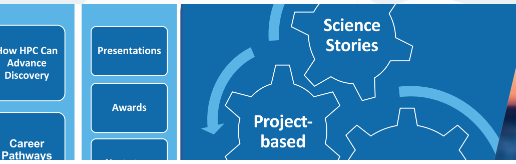
Postmaster: Send address changes to *ComputingEdge*-IEEE Membership Processing Dept., 445 Hoes Lane, Piscataway, NJ 08855. Periodicals Postage Paid at New York, New York, and at additional mailing offices. Printed in USA.

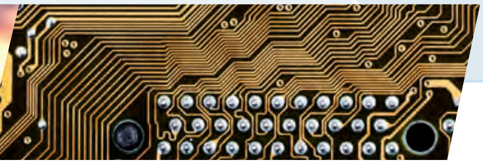
Editorial: Unless otherwise stated, bylined articles, as well as product and service descriptions, reflect the author's or firm's opinion. Inclusion in *ComputingEdge* does not necessarily constitute endorsement by the IEEE or the Computer Society. All submissions are subject to editing for style, clarity, and space.

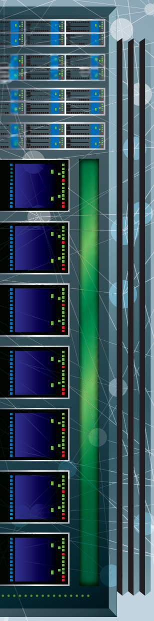
2025 IEEE Computer Society Magazine Editors in Chief

OCTOBER 2025 · VOLUME 11 · NUMBER 10

COMPUTING edge







High Performance Computing

Memory

Careers

Low Code

Departments

Magazine Roundup

The IEEE Computer Society's lineup of 12 peer-reviewed technical magazines covers cutting-edge topics ranging from software design and computer graphics to Internet computing and security, from scientific applications and machine intelligence to visualization and microchip design. Here are highlights from recent issues.

Computer

Envisioning the Next-Generation Cellular Architecture With Named Data Networking

This article, featured in the August 2025 issue of *Computer*, proposes adopting named data networking (NDN) as a foundation for future cellular networks, including 6G, to shift the focus from connection-based to data-centric communication. Securing data at the network layer, NDN reduces control plane signaling overhead and enables secure distributed deployments.

Computing in Science & Engineering

Adaptive Computing for Scale-Up Problems

Adaptive computing is an application-agnostic outer-loop framework to strategically deploy simulations and experiments to guide decision making for scale-up analysis. The framework enables the characterization and management of uncertainties associated with predictive models of complex systems when scale-up questions lead to significant model extrapolation. This January–March 2025 *Computing in Science & Engineering* article discusses applications of this framework to problems in the renewable energy space, including biofuels production, material synthesis, perovskite crystal growth, and building electrical loads.

IEEE Annals of the History of Computing

Operation Voder: AT&T, Bell Labs, and the Labor of Techno-Utopia at the 1939 New York World's Fair

This article, featured in the April–June 2025 issue of *IEEE Annals of the History of Computing*, explores the labor demands of the Voder, the electrical speech synthesis machine developed by Bell Labs for AT&T's 1939 New York World's Fair exhibit. The author argues that AT&T executives used Voder operators to normalize a new vision of technological utopia that relied heavily and conspicuously on the infrastructural labor of women. Moreover, the article highlights the previously unacknowledged engineering contributions of Voder operators in the years before the fair, writing women into the origin story of a machine that laid critical groundwork for speech recognition and voice encryption technology.

IEEE Computer Graphics and Applications

What Data Do and Do Not Represent: Visualizing the Archive of Slavery

The authors of this article, featured in the May/June 2025 issue of *IEEE Computer Graphics and Applications*, present a design report on a humanistically informed data visualization of a dataset related to the trans-Atlantic slave trade. The visualization employs a quantitative dataset of slaving voyages that took place between 1565 and 1858 and uses historical scholarship and humanistic theory to call attention to the people behind the data, as well as to what the data do not or cannot represent.

IEEE Intelligent Systems

Wet-Neuromorphic Computing: A New Paradigm for Biological Artificial Intelligence

As we delve into a life governed by artificial intelligence (AI), ongoing

research continues to discover new forms of intelligence that are efficient and closely mimic an organism's brain in terms of performance. This article, which was in the May/June 2025 issue of *IEEE Intelligent Systems,* presents a new concept termed wet-neuromorphic computing, in which biological cells or organisms are leveraged to perform computational tasks using their natural molecular functions.

Toward Carbon-Aware Data Transfers

The growing adoption of cloud, edge, and distributed computing, as well as the rise in the use of artificial intelligence/machine learning workloads, have created a significant need to measure, monitor, and reduce the carbon emissions associated with these resource-intensive tasks. One significant but often overlooked source of emissions is data transfers over wide-area networks, primarily due to the challenges in accurately measuring the carbon footprint of end-to-end network paths. The authors of this article from the March/April 2025 issue of *IEEE Internet Computing* introduce a novel mechanism to measure network carbon footprints and propose strategies for optimizing the scheduling of network-intensive tasks.

The IBM Telum II Processor

This article, featured in the May/June 2025 issue of *IEEE Micro*, presents IBM Telum II, the latest processor designed specifically for IBM Z's next-generation mainframe. Designed-for-purpose, Telum II is focused on mission-critical enterprise workloads where performance and sustainability are of the utmost importance and the demand for artificial intelligence acceleration is increasing dramatically. Innovations discussed in this article are the new on-die data processing unit for input/output acceleration, the updated cache, enhancements to the on-chip artificial intelligence accelerator, core improvements, and changes to the off-chip input/output interfaces.

A Novel Hybrid Epidemic Prediction Model Based on Cross-Modal Information

To assess the threat of the COVID-19 epidemic, forecasting the number of new cases is critical for epidemic prevention. The authors of this January–March 2025 *IEEE MultiMedia* article propose a novel cross-modal spatial-temporal epidemic prediction (CMSTEP) model for COVID-19 new-case prediction. The proposed model consists of two newly designed modules: one is a spatial-temporal sequential prediction module that captures the trend of new cases based on the historical epidemic of the target region and its related regions, the other is an intervention effect assessment module that models the NPIs' impact based on their orientations and effective durations.

Data-Driven Adaptation of Smart Grids With Hierarchical Digital Twins

Local energy communities are citizens' associations that allow efficient energy sharing and management among their members. Such organizations play a crucial role in the energy transition, and smart grids represent the core technology for their implementation. In this January–March 2025 *IEEE Pervasive Computing* article, the authors propose a framework based on hierarchical Digital Twins interconnecting the physical devices of the smart grid. By exploiting this framework, they propose an energy-sharing approach in which users of a local energy community can share the excess local batteries' capacity with each other.

IEEE SECURITY & PRIVACY

Trajectories of Piracy and Cyberbullying Across Adolescence

This article, featured in the May/June 2025 issue of *IEEE Security & Privacy*, investigates the pathways and predictors of piracy and cyberbullying among Korean adolescents using longitudinal data from 2003–2008. Findings reveal distinct trajectories and predictors for each behavior. This highlights the need for targeted interventions and challenges existing international cybercrime policies.

IEEE Software

From Code Generation to Software Testing: AI Copilot With Context-Based Retrieval-Augmented Generation

The rapid pace of large-scale software development places increasing demands on traditional testing methodologies. This article from the July/August 2025 issue of *IEEE Software* proposes a novel perspective on software testing, highlighting the transformative potential of AI-driven technologies in modern software development practices.

IT Professional

Exposing and Addressing Fake Base Station Vulnerabilities in 5G Through User Device Exploits

The rapid advancement of 5G networks introduces new security challenges, particularly with the rise of false base station (FBS) attacks. This article, featured in the May/June 2025 issue of *IT Professional*, investigates the vulnerabilities of 5G networks exploited by FBSs, which hijack communications by mimicking legitimate base stations and compromising user equipment (UE). This research provides critical insights into securing 5G networks, emphasizing the importance of adaptive defense strategies against evolving cyber threats.

Editor's Note

The Future of High-Performance Computing

Supercomputing, a type of high-performance computing (HPC), will continue to evolve over the next decade to meet the need of rising artificial intelligence/machine learning (AI/ML) use. The HPC field must adapt and grow, which includes expanding and diversifying its workforce. This issue of *ComputingEdge* discusses initiatives to bring diversity and inclusivity into HPC as well as the outlook of supercomputing, including increasing uses and adoption. The articles also explore new advancements in battery and semiconductor memory, and delve into careers in requirements engineering (RE) and video game career recruitment. The issue concludes with a discussion of developments in low-code/no-code (LCNC) platforms.

There is much on the horizon for supercomputing, from developments in AI/ML techniques to improving representation in the workforce by key communities. *Computing in Science & Engineering* article "Building a Diverse and Inclusive HPC Community for Mission-Driven Team Science" presents strategies aimed at increasing and diversifying the HPC workforce through bootcamps, internships, and a workforce development and retention group. The authors of "Predicting the Future of Supercomputing" from *Computer*, address the needs, challenges, and opportunities for supercomputing over the next decade.

The explosion of Internet of Things (IoT) devices and data-intensive applications in AI and ML is straining current memory supply and capacity. In *Computer* article "How Emerging Memories Extend Battery Life," the authors explore new nonvolatile memory types, which can help balance the tradeoffs between functionality, portability, and battery life in IoT devices. In "Semiconductor Memory Technologies: State-of-the-Art and Future Trends" from *Computer*, the authors survey the recent development of semiconductor memory technologies, which can improve capacity and speed.

Unusual career routes can lead to careers in RE, and video games can be a tool in career recruitment to attract the Gen Z workforce. *Computer* article "How to Hire a Gen Z Through Gaming" outlines a strategy for how to use any game as a recruiting tool for Gen Z. In *IEEE Software* article "My REvelation: Unveiling an Unseen Career in Requirements," the author explains how she found a career in RE.

The authors of "Citizen Development, Low-Code/No-Code Platforms, and the Evolution of Generative AI in Software Development," from *Computer*, show how an increased use of LCNC platforms combined with AI can help supplement the shortage of software developers and engineers.

EDITOR: Mary Ann Leung, mleung@shinstitute.org

This article originally appeared in
Computing in Science & Engineering
vol. 25, no. 5, 2023

Building a Diverse and Inclusive HPC Community for Mission-Driven Team Science

Lois Curfman McInnes and Paige Kinsley, *Argonne National Laboratory, Lemont, IL, 60439, USA*
Mary Ann Leung, *Sustainable Horizons Institute, Rancho Mirage, CA, 92270, USA*
Daniel Martin, *Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA*
Suzanne Parete-Koon, *Oak Ridge National Laboratory, Oak Ridge, TN, 37831, USA*
Sreeranjani (Jini) Ramprakash, *Argonne National Laboratory, Lemont, IL, 60439, USA*

The U.S. Department of Energy (DOE) has been a long-standing leader in driving advances in science and technology through advanced computing. However, DOE laboratories are currently facing urgent workforce challenges, particularly in terms of underrepresentation from key communities, including people of color, women, persons with disabilities, and first-generation scholars. This paper introduces the work carried out as part of the Exascale Computing Project (ECP) Broadening Participation Initiative, which aims to address workforce challenges through a lens that considers the distinct needs and culture of high-performance computing (HPC). The work focuses on three main efforts: hosting Intro to HPC Bootcamps, expanding the Sustainable Research Pathways (SRP) internship and workforce development program, and establishing an HPC Workforce Development and Retention Action Group. The paper also highlights various workforce efforts throughout the computational science community and explores opportunities for future work aimed at broadening participation in HPC.

The U.S. Department of Energy (DOE) is a long-standing leader in scientific discovery enabled through high-performance computing (HPC). Associated with 118 Nobel Prize winners, the 17 DOE national laboratories conduct a wide array of basic and applied science research, with emphasis on solving big problems through mission-driven team science. DOE's investments have pushed the growth of computational and data-enabled science and engineering as a foundation of scientific and technological progress in conjunction with theory and experimentation.[1] Computational science—at the intersection of mathematics and statistics, computer science, and core disciplines of science and engineering—is revolutionizing not only the traditional physical sciences, but also life sciences, social sciences, humanities, business, finance, and even government policy.

1521-9615 © 2024 IEEE
Digital Object Identifier 10.1109/MCSE.2023.3348943
Date of current version 12 April 2024.

BUILDING THE WORKFORCE TO TACKLE BIG PROBLEMS THROUGH HPC TEAM SCIENCE

As we tackle next-generation challenges and problems otherwise intractable—bridging scales and domains through new multiscale and multiphysics algorithms that exploit advanced computing architectures, incorporating complex workflows that couple modeling/simulation and experimental/observational data, leveraging artificial intelligence/machine learning (AI/ML) tools for enhanced insight, and working toward greater scientific reproducibility—we face a new era of complexity.

Past success has relied on developing a highly inter- and multidisciplinary workforce and culture that fosters cross-disciplinary communication and not only exploits but also celebrates the unique expertise of each field. The combined expertise of diverse teams is increasingly essential, including applied mathematicians, computer scientists, domain scientists, and research software engineers,[a] along with project coordinators, social scientists, and more.[2] Moreover, various studies have shown that diverse organizations and groups are more creative, innovative, and productive.[3,4]

HPC Workforce Challenges

DOE national laboratories, like many other scientific research organizations, face growing needs and challenges in recruiting and retaining a skilled workforce in the computing sciences.[5] HPC has additional constraints stemming from its reliance on a workforce versed not only in advanced computing but also in multi- and interdisciplinary science and engineering domains, which also face challenges in recruiting and retaining underrepresented populations.[b] Government and academic sectors face fierce competition for talent attracted to lucrative industrial workplace benefits. Moreover, the changing U.S. demographics and higher attrition rates among people from underrepresented groups present additional challenges.

Cultivating the HPC workforce appears to be an over-constrained problem: growing needs, higher competition, changing workforce demographic profiles, and higher attrition rates in demographic groups currently underrepresented in HPC, but growing in the general workforce population. Moreover, while HPC has successfully cultivated a technically diverse workforce and many successful recruitment models exist, widespread reliance on existing social and professional networks has largely resulted in a homogeneous workforce.[c] The challenge is not only to develop new approaches to broaden the reach but also to change longstanding recruitment, onboarding, and retention practices to create and sustain an inclusive and diverse HPC workforce.

Advancing the HPC Workforce

Addressing these workforce challenges requires broad community collaboration to change the culture and demographic profile of computational science. Impactful DOE-wide programs such as SULI,[d] GEM,[e] VFP,[f] CCI,[g] and activities[h] in the wider computing community[7,8] are making headway. Likewise, events such as Advanced Computing for Social Change,[9] The Pipeline Workshop,[i] and Scaling HPC Education[j] are pioneering innovative formats to engage underrepresented students in HPC. In addition, various communities are exploring strategies to improve HPC education and training; for example, a working group[10] made recommendations for overcoming key challenges in undergraduate-level education in computing and HPC, including building an HPC educator community and developing and providing inexpensive HPC hardware as teaching tools. Meanwhile, laboratory-specific initiatives are addressing challenges in workforce and training, capitalizing on each lab's unique perspectives, culture, and regional connections to underrepresented populations.

ECP BROADENING PARTICIPATION INITIATIVE

The DOE Exascale Computing Project (ECP)[k] is a research, development, and deployment project spanning multiple national labs as well as academic and private institutions. Beginning in 2016, ECP has engaged 1000 researchers over seven years on the development of an integrated scientific computing software stack for use on exascale supercomputers (capable of executing 10^{18} operations per second) and the demonstration of new and faster capabilities in a wide variety of applications in chemistry, materials, energy, Earth and space sciences, data analytics, optimization, AI, and

[a] https://us-rse.org, https://society-rse.org

[b] As discussed in a 2023 NSF report, https://ncses.nsf.gov/pubs/nsf23315/, women and racial and ethnic minorities are underrepresented in U.S. science and engineering programs.

[c] Demographic data for DOE national laboratories provides workforce insights, https://nationallabs.org/staff/diversity. According to a 2021 study of nine HPC and HPC-related conferences, women represent only 10% of all HPC authors.[6] A 2018 *Wired* article discusses reasons for underrepresentation of women and minorities in technology fields, https://www.wired.com/story/computer-science-graduates-diversity.

[d] DOE Science Undergraduate Laboratory Internships (SULI, https://science.osti.gov/wdts/suli) encourage undergraduate students to pursue STEM careers by providing research experiences at DOE laboratories.

[e] The GEM Fellowship Program (https://gemfellowship.org) seeks to recruit high-quality underrepresented students looking to pursue degrees in applied science and engineering.

[f] The DOE Visiting Faculty Program, (VFP, https://science.osti.gov/wdts/vfp) seeks to increase the research competitiveness of faculty members and their students at institutions historically underrepresented in the research community.

[g] The DOE Community College Internships Program (CCI, https://science.osti.gov/wdts/cci) seeks to encourage community college students to enter technical careers relevant to the DOE mission.

[h] Best Practices for Diversity and Inclusion in STEM Education and Research: A Guide by and for Federal Agencies, National Science and Technology Council, https://www.whitehouse.gov/wp-content/uploads/2021/09/091621-Best-Practices-for-Diversity-Inclusion-in-STEM.pdf.

[i] https://cra.org/cra-wp/events/pipeline-workshop-diversifying-hpc-workforce/

[j] https://supercloud.mit.edu/scaling-hpc-education

[k] https://exascaleproject.org

national security.[11] Advanced software technologies—including programming models and runtimes, mathematical libraries, data and visualization packages, and development tools that constitute the Extreme-Scale Scientific Software Stack (E4S)[l]—form a community software ecosystem that underpins ECP applications and is unlocking the potential of advanced computing across all scales.[12]

The technical breadth and sustained multilab collaboration of ECP have provided a unique and compelling opportunity for the DOE HPC community to address workforce challenges through a lens that focuses on the distinct needs and culture of DOE HPC, with its emphasis on mission-driven team science.[m] Consequently, in August 2021 the ECP Broadening Participation Task Force was established, with members representing eight DOE national laboratories—Argonne (ANL), Brookhaven (BNL), Lawrence Berkeley (LBNL), Lawrence Livermore (LLNL), Los Alamos (LANL), Oak Ridge (ORNL), Pacific Northwest (PNNL), and Sandia (SNL)—as well as the DOE Office of Science computing facilities: Argonne Leadership Computing Facility (ALCF), National Energy Research Scientific Computing Center (NERSC), and Oak Ridge Leadership Computing Facility (OLCF). After clarifying the most urgent workforce challenges in the DOE computing sciences and surveying relevant ongoing work, the task force leveraged ECP's unique position as a broad effort spanning the DOE computational research ecosystem to launch the ECP Broadening Participation Initiative.[n] The initiative embodies a collaboration among ECP investigators, facilities staff, and education and workforce professionals. On a path toward a post-ECP role, it has expanded to invite participation from all lab staff in the DOE computing sciences.

As shown in Figure 1 and discussed in the following sections, the ECP Broadening Participation Initiative features three complementary thrusts: 1) launching the Intro to HPC Bootcamp, an immersive program designed to engage students in energy justice using project-based pedagogy and real-life science stories to teach foundational skills in HPC, scalable AI, and analytics, while exposing students to the excitement of DOE mission-driven team science; 2) expanding the Sustainable Research Pathways (SRP) internship and workforce development program as a multilab cohort of students from underrepresented groups (and faculty working with them), who collaborate with DOE lab staff on world-class R&D projects; and 3) establishing an HPC Workforce Development and Retention Action Group to foster a supportive and inclusive culture in DOE labs and communities.

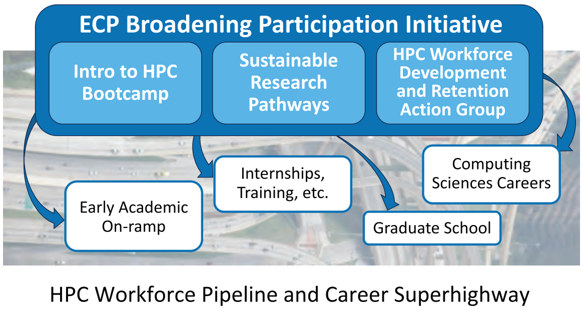

ECP Broadening Participation Initiative
Intro to HPC Bootcamp
Sustainable Research Pathways
HPC Workforce Development and Retention Action Group
Internships, Training, etc.
Computing Sciences Careers
Early Academic On-ramp
Graduate School
HPC Workforce Pipeline and Career Superhighway

FIGURE 1. The ECP Broadening Participation Initiative supports the full HPC workforce pipeline, from onramps to career retention.

These three thrusts provide paths for student engagement and retention at multiple points of the HPC workforce timeline, increasing access to and enhancing the DOE HPC community. If you imagine the HPC academic and career pathway as a superhighway, as depicted in Figure 1, we envision the three thrusts as onramps at different points, each meeting students (and lab staff) where they are, providing support and preparation, and offering access at appropriate points in an individual's journey. Through these three thrusts, the ECP Broadening Participation Initiative supports the full life cycle of the academic and career pipeline.

INTRODUCTION TO HPC BOOTCAMP

The Introduction to HPC Bootcamp serves as the first entry point onto the HPC career superhighway for students early in their academic careers. The bootcamp is designed to engage students who may not know how or why HPC could help them accomplish their academic and scientific goals, while also preparing them for internships and inspiring them to continue their studies in graduate school.

To this end, the first Intro to HPC Bootcamp[o] was developed and organized by the advanced computing facilities at ANL, LBNL, and ORNL in collaboration with Sustainable Horizons Institute, taking place in August 2023 at LBNL.[13] The bootcamp brought together 60 students to work in project groups supported by 14 trainers made up of national lab staff and academic partners and 10 peer mentors. Each group of students

[l] https://e4s-project.github.io
[m] R. Giles et al., "Transforming ASCR after ECP," https://science.osti.gov/-/media/ascr/ascac/pdf/meetings/202004/Transition_Report_202004-ASCAC.pdf.
[n] https://www.exascaleproject.org/hpc-workforce
[o] https://shinstitute.org/intro-to-hpc-bootcamp

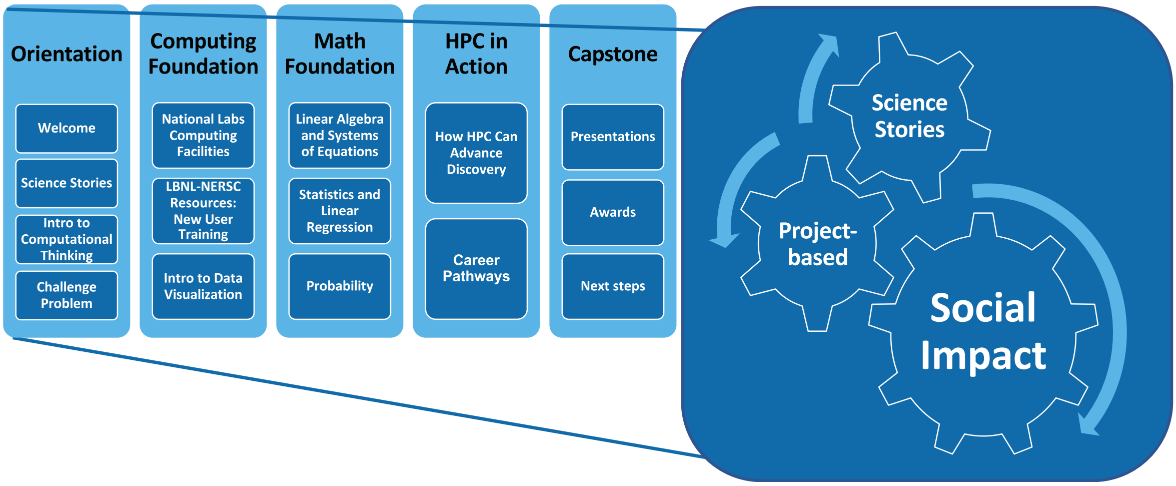

Orientation
Welcome
Science Stories
Intro to Computational Thinking
Challenge Problem
Computing Foundation
National Labs Computing Facilities
LBNL-NERSC Resources: New User Training
Intro to Data Visualization
Math Foundation
Linear Algebra and Systems of Equations
Statistics and Linear Regression
Probability
HPC in Action
How HPC Can Advance Discovery
Career Pathways
Capstone
Presentations
Awards
Next steps
Science Stories
Project-based
Social Impact

FIGURE 2. The framework of the Intro to HPC Bootcamp focuses on solving problems with mission-driven social impact to engage new communities in HPC.

explored one of seven energy justice projects developed from DOE science that examined issues related to sustainable energy usage and alternatives, social impact of climate risk and resilience, and energy equity in the United States. The bootcamp raised awareness of the power and benefits of HPC, engaging students from historically underrepresented groups in HPC by fostering a sense of belonging and exposing students to opportunities in HPC, especially at DOE labs.

The bootcamp utilized culturally relevant pedagogy and project-based learning to engage students in addressing social impact questions related to energy justice. As shown in Figure 2, the five-day bootcamp began by building community and a friendly learning environment, motivating participants through socially relevant scientific problems, while exposing students to foundational concepts in computing, HPC, mathematics, and analysis. Learners worked throughout the week on their projects, culminating in a presentation on the final day. Throughout the bootcamp, students had opportunities to hear from lab staff about their paths to the national labs and HPC careers.

To ensure an engaging and inclusive bootcamp, we included a diverse set of organizers, trainers, and mentors to develop and facilitate the program. The team had expertise in computational science, advanced computing, energy justice, diversity, education, workforce development, and program evaluation. Collaborators came from multiple DOE labs, Sustainable Horizons Institute, the DOE Office of Economic Impact and Diversity, and academia. Alongside the bootcamp trainers, peer mentors provided guidance on technical concepts, collaboration, workshop expectations, and presentations.

A pivotal step in developing the bootcamp was a Train the Trainers workshop, essential for building effective team collaboration and introducing the bootcamp concept to the DOE lab staff who would serve as materials developers and trainers. During the workshop, the team considered strategies for modifying existing HPC training materials through the lens of engaging new communities and project-based learning.

The bootcamp application was designed to lower the barriers for students historically underrepresented in HPC, asking why students wanted to attend the bootcamp without requiring a letter of recommendation or curriculum vita. Of the over 300 students who applied, 60 were chosen to attend; all had some experience with computing but little or no background in HPC. Nearly 80% were undergraduates, along with a mix of master's, Ph.D., and community college students. As shown in Figure 3, the 60 students were diverse racially, with African Americans/Black representing the largest group at 28%, followed by Asian and multiple race (20%) and Caucasian (18%). Nearly all participants (98%) identified as a member of at least one underrepresented group, including 48% first-generation scholars and 57% women participants. For these 60 students, travel, lodging, and food were covered, and a US$500 stipend was provided at the end of the bootcamp to ensure all students, regardless of financial context, would be able to attend.

Preliminary feedback about the bootcamp has been positive. Of the 54 students who responded to a post-bootcamp survey, 85% said they are interested or very interested in a career in HPC, and 90% said they are interested or very interested in a career at a DOE

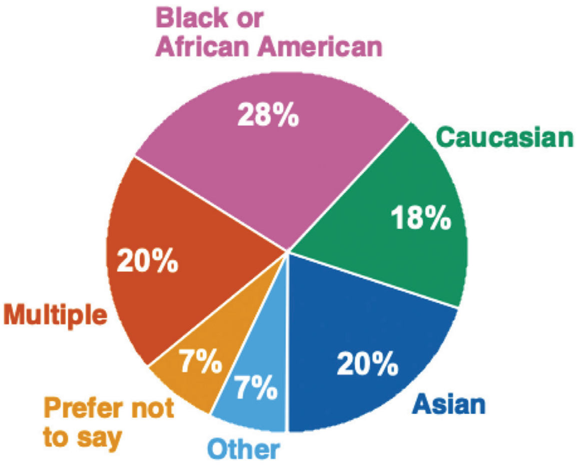


FIGURE 3. Breakdown of demographics of participants of the Intro to HPC Bootcamp in August 2023.

national lab. In responses to the post-bootcamp survey, many students expressed changes in potential future plans because of their experience at the bootcamp: "[The Intro to HPC Bootcamp] exposed me to work and life at a national lab and motivated me to consider such work … ." "If not at a national lab, it has definitely motivated me to look into non-industry jobs that more greatly benefit the public good." "I think I would find that incredibly meaningful and fulfilling, and that is very important to me in a career and something I've been searching for … ."

Building on bootcamp successes, we plan to increase access to introductory HPC materials by building in adaptability and customization. Potential paths to increase sustainability and access to the program are through faculty partnerships, asynchronous bootcamp components, providing local offerings, and creating curriculum components to progressively build HPC skills for more advanced learners. By reaching more participants, we are working to build a sustainable pipeline of talent for DOE national labs, preparing students for internship opportunities and providing them with tools to succeed in the next steps of HPC careers.

SUSTAINABLE RESEARCH PATHWAYS

Sustainable Research Pathways (SRP)[p] is a comprehensive workforce development program designed to increase the participation of underrepresented groups and institutions in research and development at DOE national labs and to address pressing needs in the advanced scientific computing workforce. SRP serves as an on-ramp for undergraduate and graduate students with some experience and interest in computing or computational science and engineering. Started as a partnership between Sustainable Horizons Institute and LBNL in 2015,[14] SRP expanded beyond LBNL in 2022 as part of the ECP Broadening Participation Initiative. The 2022 SRP cohort was comprised of a highly diverse group of faculty, students of faculty, and independent students who collaborated with staff at 10 DOE national labs on topics involving advanced software technologies and scientific applications. In 2023, the program further expanded through a partnership with ECP and seven labs (ANL, BNL, LBNL, LLNL, LANL, ORNL, SNL) from the Computational Research Leadership Council. During summer 2023, SRP facilitated nearly 200 faculty and student collaborations at 10 DOE labs, and work is underway to prepare for the summer 2024 cohort.

Figure 4 illustrates the program components and timeline. SRP begins with extensive recruitment of faculty, students, and DOE staff members for participation in an interactive virtual SRP Matching Workshop, where participants engage in successively more focused interactions, providing them with the opportunity to explore common interests and potential collaborative summer projects. At the conclusion of the workshop, participants indicate their matching preferences, which are used to create two-way matches (requiring that both parties have expressed interest in working together). Following the matching process, the resulting teams develop brief project plans/proposals for the summer experience, and then funded teams are onboarded at their assigned DOE lab. In addition to pursuing their summer research projects, participants attend local seminars and social events at their host laboratory, as well as virtual cohort-wide SRP activities that help build a multilab SRP community. Volunteer committees provide participants with professional development and leadership opportunities; recent activities have included game nights, wellness events, and seminars on work/life balance. After the summer internship, the Catalyzing Ubiquitous Learning Through InoVaTive and Engaging (CULTIVATE) Conversations program serves to maintain contact with the cohort with the aim of helping develop inclusive workforce ecosystems through facilitated conversations. Participants often showcase their SRP accomplishments at scientific conferences.

SRP faculty alumni accomplishments include DOE Early Career, ASCR-RENEW, and ASCR FAIR awards and nomination for the Presidential Awards for Excellence in Science, Mathematics, and Engineering Mentoring. Student participants have been hired full-time at LBNL, LLNL, and PNNL; decided to pursue graduate

[p] https://shinstitute.org/sustainable-research-pathways-2024-workshop

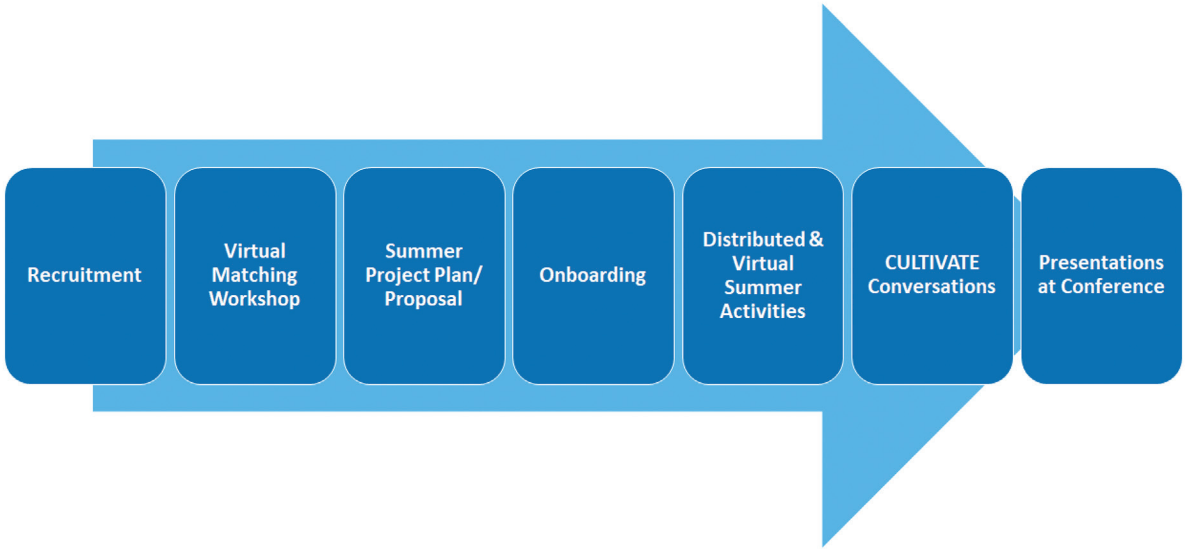

Recruitment
Virtual Matching Workshop
Summer Project Plan/ Proposal
Onboarding
Distributed & Virtual Summer Activities
CULTIVATE Conversations
Presentations at Conference

FIGURE 4. Components of the SRP program engage students to promote multifaceted learning and community building across a broad timeline before, during, and after the summer internship experience.

degrees; won best poster awards; been awarded the DOE Computational Science and Stockpile Stewardship Graduate Fellowships and Fulbright scholarships; presented their research at prestigious venues such as the Supercomputing and SIAM CSE conferences; and participated in programs such as the Argonne Training Program on Extreme-Scale Computing[q] and the ORNL Artificial Intelligence Workshop.

SRP forms meaningful and lasting connections between faculty, students, and DOE lab staff, while helping to build inclusive HPC research and development ecosystems. Community building activities start at the virtual SRP Matching Workshop and continue during the summer experience and beyond through CULTIVATE Conversations.

HPC WORKFORCE DEVELOPMENT AND RETENTION ACTION GROUP

While the Intro to HPC Bootcamp and Sustainable Research Pathways prepare students to enter the HPC career superhighway, we recognize the vital importance of cultivating an inclusive ecosystem, not only to welcome and retain them, but also to foster an environment where they can thrive, meet their potential, and express their full selves. As shown in Figure 1, we complement work to engage students in the HPC workforce pipeline with activities aimed at creating a culture of inclusion. The HPC Workforce Development and Retention (HPC-WDR) Action Group facilitates collaboration among DOE national laboratories and their associated computing communities to share knowledge and insights aimed at creating a diverse, equitable, and inclusive workforce for HPC. This effort focuses on building a community and gathering proven tools and best practices. Initially, representatives from ten national laboratories convened regularly to exchange ideas and develop recommendations and strategies for building supportive workforce cultures. The first two activities undertaken by HPC-WDR have been organizing a quarterly webinar series on HPC workforce topics[r] and establishing a dedicated website that focuses on fostering a diverse and inclusive HPC workforce culture as well as addressing workforce retention.

The webinars have explored topics such as effective mentoring practices and the significance of embracing diversity for inclusion. Since their inception in May 2022, seven webinars in the series have garnered participation from 672 individuals, representing 10 national laboratories, 38 universities, and 22 businesses, with speakers drawn from the scientific computing community; recordings enable even broader reach.

The HPC-WDR website[s] serves as a repository for webinar recordings and provides announcements of computing workforce events. Moreover, the website houses a growing collection of best practices on HPC workforce issues, often presented in blog posts. For

[q] https://extremecomputingtraining.anl.gov

[r] https://www.exascaleproject.org/workforce-development-seminar-series

[s] https://hpc-workforce-development-and-retention.github.io/hpc-wdr

instance, one blog discusses the adoption of "inclusive minutes" during team meetings.[t] In this practice, teams allocate a minute during their meetings to exchange insights on integrating inclusive and culturally aware practices into their work areas, with a goal of improving communication and fostering mutual respect. Recognizing that changing workplace culture is a complex problem with long timescales, our ongoing objective is to maintain the website as a living community resource and to steward and advance our community's presence. We intend to continue hosting webinars and workforce community meetings, recognizing these activities as vital for assisting the DOE labs' computing community in identifying and implementing best practices in workforce development and retention. Our initial efforts focused on establishing the working group. With a gratifying response from webinar attendees, our next focus is to develop a mixture of qualitative and quantitative methods for capturing impacts and to establish data collection points to measure changes over time, thus motivating further work.

FUTURE DIRECTIONS

The ECP Broadening Participation Initiative has established a strong foundation for collaborating and innovating as a multilab community to address challenges in the complete life cycle of the DOE HPC workforce. Early successes include an overwhelming response to the call for participation in the innovative energy justice project-based HPC bootcamp; phenomenal SRP growth from a single lab to a multilab initiative that fostered nearly 200 collaborations across 10 labs in the summer of 2023; and the establishment of a highly collaborative group of laboratory, academic, and industrial HPC professionals who have shared best practices, established a repository of materials, and facilitated webinars attended by nearly 700 people.

The three thrusts—the Intro to HPC Bootcamp, Sustainable Research Pathways, and the HPC Workforce Development and Retention Action Group—provide different entry points onto what we imagine as the fast-paced, exciting, multi-, and interdisciplinary HPC career superhighway. Through this multipronged approach, we can attract students of varied backgrounds, experience levels, and interests, wherever they are in their journeys. We further invite them to join us as we work toward an HPC community where everyone can reach their potential, be their full selves, and contribute to a more innovative mission-driven team science enterprise.

Together, as we plan for the next phases of work to broaden the participation of underrepresented groups, we are working to realize a sustainable strategy to recruit and retain a diverse HPC workforce by fostering a supportive and inclusive culture within the computing sciences at DOE national laboratories. The exciting and complex era of next-generation computational science demands a multidisciplinary workforce whose members provide a diversity of technical expertise *and* are fully representative of our whole population. This diversity across many axes will inspire innovation, provide new perspectives, and enable us to tackle big problems through HPC team science.

ACKNOWLEDGMENTS

This work was supported by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy, Office of Science, and the National Nuclear Security Administration. We sincerely thank all those who have contributed to the ECP Broadening Participation Initiative for their passion and commitment to advance the DOE workforce in computing sciences. We especially thank members of the ECP Task Force on Broadening Participation, who conceived of and launched the three complementary thrusts; members of the HPC Workforce Development and Retention Action Group, who meet regularly for discussion and actions to advance issues in workforce inclusion; mentors in the Sustainable Research Pathways Program, who host participants from underrepresented groups for summer internships; the team who planned and led the Intro to HPC Bootcamp, including trainers at ALCF, NERSC and OLCF and university partners; members of the Computational Research Leadership Council, who provide support for multilab partnership; ECP leadership; and the DOE Office of Advanced Scientific Computing Research and the National Nuclear Security Administration. This work used resources of the Argonne Leadership Computing Facility, the Oak Ridge Leadership Computing Facility, and the National Energy Research Scientific Computing Center, which are DOE Office of Science User Facilities supported under Contract DE-AC02-06CH11357, Contract DE-AC05-00OR22725, and Contract DE-AC02-05CH11231, respectively.

[t] https://hpc-workforce-development-and-retention.github.io/hpc-wdr/jekyll/update/2023/04/08/inclusive-minute.html

this manuscript, or allow others to do so, for U.S. government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (https://www.energy.gov/doe-public-access-plan).

REFERENCES

1. B. Hendrickson et al., "ASCR@40: Highlights and impacts of ASCR's programs," U.S. Department of Energy, Washington, DC, USA, 2020. [Online]. Available: https://www.osti.gov/biblio/1631812
2. M. A. Heroux et al., "Basic research needs in the science of scientific software development and use: Investment in software is investment in science," U.S. Department of Energy, Washington, DC, USA, 2023. [Online]. Available: https://www.osti.gov/servlets/purl/1846009
3. D. Rock and H. Grant, "Why diverse teams are smarter," *Harvard Bus. Rev.*, Nov. 2016. [Online]. Available: https://hbr.org/2016/11/why-diverse-teams-are-smarter
4. D. V. Hunt, D. Layton, and S. Prince, "Why diversity matters," McKinsey and Company, Chicago, IL, USA, Tech. Rep., 2015. [Online]. Available: https://www.mckinsey.com/capabilities/people-and-organizational-performance/our-insights/why-diversity-matters
5. B. Chapman et al., "DOE advanced scientific advisory committee (ASCAC): Workforce subcommittee letter," U.S. Department of Energy, Washington, DC, USA, 2014. [Online]. Available: https://www.osti.gov/servlets/purl/1222711
6. E. Frachtenberg and R. D. Kaner, "Representation of women in HPC conferences," in *Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal. (SC)*, 2021, pp. 1–14, doi: 10.1145/3458817.3476164.
7. T. Whitney and V. Taylor, "Increasing women and underrepresented minorities in computing: The landscape and what you can do," *Computer*, vol. 51, no. 10, pp. 24–31, Oct. 2018, doi: 10.1109/MC.2018.3971359.
8. E. Hammonds, V. Taylor, and R. Hutton, Eds., *Transforming Trajectories for Women of Color in Tech*. Washington, DC, USA: The National Academies Press, 2022.
9. K. Gaither et al., "Advanced computing for social change: Educating and engaging our students to compete in a changing workforce," in *Proc. Pract. Experience Adv. Res. Comput. Sustainability, Success Impact (PEARC)*, Jul. 2017, pp. 1–4, doi: 10.1145/3093338.3093391.
10. R. K. Raj et al., "High performance computing education: Current challenges and future directions," in *Proc. Working Group Rep. Innov. Technol. Comput. Sci. Educ. (ITiCSE-WGR)*, New York, NY, USA: Association for Computing Machinery, 2020, pp. 51–74, doi: 10.1145/3437800.3439203.
11. F. Alexander et al., "Exascale applications: Skin in the game," *Philos. Trans. Roy. Soc. A, Math., Physical Eng. Sci.*, vol. 378, no. 2166, Mar. 2020, Art. no. 20190056, doi: 10.1098/rsta.2019.0056.
12. L. C. McInnes, M. A. Heroux, E. W. Draeger, A. Siegel, S. Coghlan, and K. Antypas, "How community software ecosystems can unlock the potential of exascale computing," *Nature Comput. Sci.*, vol. 1, no. 2, pp. 92–94, 2021, doi: 10.1038/s43588-021-00033-y.
13. M. A. Leung et al., 2023, "Intro to HPC Bootcamp: Engaging new communities through energy justice projects," Figshare, doi: 10.6084/m9.figshare.24168675.
14. M. A. Leung, S. Crivelli, and D. Brown, "Sustainable research pathways: Building connections across communities to diversify the national laboratory workforce," Collaborative Network for Engineering and Computing Diversity (CoNECD), Washington, DC, USA, 2019. [Online]. Available: https://peer.asee.org/sustainable-research-pathways-collaborations-across-communities-to-diversify-the-national-laboratory-workforce

LOIS CURFMAN MCINNES is a senior computational scientist and Argonne Distinguished Fellow at Argonne National Laboratory, Lemont, IL, 60439, USA. Contact her at mcinnes@anl.gov.

PAIGE KINSLEY is the education outreach lead at the Argonne Leadership Computing Facility, Argonne National Laboratory, Lemont, IL, 60439, USA. Contact her at pkinsley@anl.gov.

MARY ANN LEUNG is founder and president of Sustainable Horizons Institute, Rancho Mirage, CA, 92270, USA. Contact her at mleung@shinstitute.org.

DANIEL MARTIN is a staff scientist in the Applied Numerical Algorithms Group, Lawrence Berkeley Laboratory, Berkeley, CA, 94720, USA. Contact him at dfmartin@lbl.gov.

SUZANNE PARETE-KOON is a high-performance computing engineer at Oak Ridge National Laboratory, Oak Ridge, TN, 37831, USA. Contact her at paretekoonst@ornl.gov.

SREERANJANI (JINI) RAMPRAKASH is the deputy division director at the Leadership Computing Facility, Argonne National Laboratory, Lemont, IL, 60439, USA. Contact her at ramprakash@anl.gov.

EDITOR: Dejan Milojicic, Hewlett Packard Labs, dejan.milojicic@hpe.com

This article originally appeared in
Computer
vol. 58, no. 7, 2025

DEPARTMENT: PREDICTIONS

Predicting the Future of Supercomputing

Scott Atchley, *Oak Ridge National Laboratory*
Rosa M. Badia, *Barcelona Supercomputing Center*
Bronis R. de Supinski, *Lawrence Livermore National Laboratory*
Joshua Fryman, *Intel*
Dieter Kranzlmüller, *Leibniz Supercomputing Centre*
Srilatha Manne, *Advanced Micro Devices, Inc.*
Pekka Manninen, *IT Center for Science*
Satoshi Matsuoka, *RIKEN*
Dejan Milojicic, *Hewlett Packard Labs*
Galen Shipman, *Los Alamos National Laboratory*
Eric Van Hensbergen, *Arm*
Robert W. Wisniewski, *Hewlett Packard Enterprise*

The need to solve high-complexity problems using large-scale tightly coupled computing (that is, supercomputing) continues to grow. In this article, we address the needs, challenges, and opportunities for supercomputing over the next decade.

Supercomputing, which involves the use of the highest-performance computing resources available at a given time, has recently seen broader adoption as it is essential for training generative artificial intelligence/machine learning (AI/ML) models. These AI use cases are in addition to the traditional modeling and simulation (modsim) workloads that continue to drive high use at traditional supercomputing centers.

Supercomputing centers are increasingly adopting AI/ML techniques into modsim workloads. This article by leaders from those centers, as well as within the industry, explores the trends and directions that will shape future supercomputers, driven largely by that convergence of modsim and AI/ML techniques. This article extends the predictions of several recent articles that explored the future of supercomputing.[1,2,3,4,5,6]

INCREASING USES, INCREASING ADOPTION

As we consider the future of supercomputing, we see several factors that will drive changes to the workloads that are run on supercomputers. These changes will continue to broaden the adoption of supercomputing and will affect the technology used to build supercomputers. In this section and the following one, we describe our expectations for future supercomputing workloads and discuss the technologies that will shape their evolution.

While we expect supercomputing workloads to be augmented with new workloads (for example, AI), we expect that traditional supercomputing workloads will remain a significant use case. These traditional workloads serve a wide range of purposes, from advancing science to deepening our understanding of the universe

Digital Object Identifier 10.1109/MC.2025.3567232
Date of current version: 27 June 2025

in which we live, addressing humanity's needs in the modern world, to protecting the national interests of governments that deploy such systems. Nonetheless, we expect these traditional workloads to incorporate new algorithmic techniques, starting with the use of AI/ML models, as has already begun.[7,8,9] The adoption of AI/ML techniques includes their use to guide the simulated configurations in ensemble calculations but also their use to accelerate expensive calculations of models of physics and biological phenomena.

With the end of Dennard scaling and the slowing of Moore's law, the automatic increase in performance at constant cost and power is over. Modsim practitioners are faced with modest gains in performance with incremental architecture changes. Future gains are largely coming from the increase in silicon within the package. While providing needed performance boosts, it comes with higher power and higher costs for both the additional silicon and the integration to stitch together several chiplets. When viewed as performance per watt (for example, if a facility has a fixed power budget), then the gains are still modest.

At the same time, the explosive growth in AI, both training and inference, has driven silicon vendors to tailor their products to this lucrative market. It is not clear, however, that modsim can take advantage of lower precision. Some apps will be able to use FP32 for some of their data structures (but not necessarily all) and see benefits compared to lazily promoting everything to FP64. It is not clear if apps will be able to use FP16 for modsim unless it is using AI inferencing in lieu of a component in a multiphysics application, emulation, or iterative refinement. To use ML inferencing, there needs to be an already-trained model. There is a lot of research interest in determining when/if modsim applications can exploit lower precision, which is becoming much more plentiful. There are efforts to see which, if any, apps can use lower precision directly, use lower precision via AI methods, use lower precision via iterative refinement, or use lower precision via emulation. Some apps may be able to do so, while others will not.

The beauty of the General Matrix-Matrix Multiplication (GEMM) emulation methods (that is, Ozaki methods[10]) is that precision is finer-grained than with hardware. Hardware is limited to powers of two (for example, FP64, FP32, and FP16), while Ozaki can provide any multiple of four bits (for example, FP40, FP48, and FP56) to provide just enough precision to converge on a valid solution without providing "too much." While Ozaki's scheme can outperform native cuBLAS in some cases, the downsides to emulation are 1) it can only emulate GEMM (that is, matrix-matrix) instructions but not vector instructions, and 2) it consumes 30–50% of the available memory, thus reducing the solvable problem size. If memory were cheap and plentiful, the latter would not be an issue, but supercomputer users want the fastest memory available. Today, that is high-bandwidth memory, and it is neither cheap nor plentiful. Recently, systems used for AI/ML training have been cast as competitors to supercomputers.

Rather than competitors, the authors view both modsim and AI as having overlapping needs for supercomputer design, except for precision. However, the systems that provide AI/ML capability are best viewed as supercomputers themselves and reflect that AI/ML training has emerged as an important workload for supercomputers. As we look toward the future, not only do we expect that AI/ML training will remain a critical supercomputing workload, but we anticipate that additional new workloads will emerge. We expect that domains that have begun to use supercomputers more extensively due to the success of large-scale AI/ML models, such as finance and retail, will identify new mechanisms to exploit the computational capability available and expand the use of AI/ML in their domain.

The convergence of cloud computing and supercomputing has long been expected. However, this convergence has not fully materialized yet, in part due to the requirements of traditional tightly coupled parallel modsim workloads. Nonetheless, cloud providers continue providing more high performance computing (HPC) capability, and cloud computing continues to be a viable economic and technical alternative for embarrassingly parallel workloads and, as of recently, for AI/ML workloads. They are also suitable for offloading or bursting small-scale experiments and development.

Addressing humanity's needs, such as weather forecasting and biomedical research, continues to be an important target of supercomputing. These applications include energy needs and its production using nuclear fission near term and fusion long term—but also for fossil fuels and importantly, carbon and water management. Another use is for new materials, particularly for the continued advancement of technology beyond silicon CMOS device scaling. Yet another use case is mitigating and adapting to climate change, including utilizing digital twins.

Digital twins are virtual representations of physical artifacts, systems, or processes with collected real-time information. They enable monitoring, simulation, and prediction of those physical artifacts. Digital twins often use supercomputers directly in a variety of vertical applications and services (for example, for structural analysis, Earth monitoring, manufacturing, and operations) as well as exploit them peripherally (for example, for monitoring, optimizing operation, anomaly detection, or what-if-analysis). Digital twins are used in areas such as the transportation industry, data centers,[11,12] and even Earth.[13]

Another important use case of traditional supercomputing is helping drive new scientific breakthroughs [that is, helping answer the big questions, for example, performing computation for the follow-on to Laser Interferometer, Gravitational-Wave Observatory (LIGO) or Laser Interferometer Space Antenna (LISA) that will enable sensing of gravitational wavelengths populated by a rich diversity in astrophysical phenomena that are of deep interest to astronomers and astrophysicists]. After a discussion on how the use and adoption of supercomputing evolved, we will next explore how technology evolution impacts workloads.

EVOLVING TECHNOLOGIES AND WORKLOADS

Future supercomputing workloads will reflect recent and anticipated future technological and industry developments. These trends include not only the adoption of AI/ML to serve edge computing and other end-user applications but also productivity enhancements, such as those driving broad consumer adoption of cloud-based computing. Further, architectural and device-level advancements will continue to motivate new supercomputing application enhancements. This section provides a high-level description of these two influences on future supercomputers. We begin by describing the workloads.

- New applications are continuing to demand more computational capability, including bioengineering, climate modeling, national security, fusion energy, and many others.
- HPC and AI will continue to converge and thereby demand more AI-ready infrastructure.
- Large language models and other models have captured the public imagination, and they open new opportunities in supercomputing.
- Physics-informed neural networks and other models, possibly integrated into traditional modsim applications, enable the faster exploration of design spaces.
- Some workloads are increasing performance by leveraging mixed precision computation, while others are leveraging multitenancy to increase performance.
- Application demand for scale-up networking, including Ultra Accelerator Link (UALink), will continue to increase per-device bandwidth and the number of directly connected scale-up devices, blurring the boundary between scale-up and scale-out infrastructure.

In the last couple of years, advancements in AI, specifically in generative AI applications, have dramatically influenced private industry toward building large-scale computing infrastructure. Even though these infrastructures are driven by AI requirements, they are becoming increasingly HPC ready. AI and HPC are making significant strides toward convergence, and this development is a major disruptor. We predict the forthcoming technological changes.

- Accelerators, from traditional (for example, compression and crypto) to ones focused on

AI (for example, Cerebras, NextSilicon, and SambaNova) to upcoming (for example, neuromorphic and quantum), will address specialized but important demands, and some are already being incorporated into existing supercomputers. 2.5D and 3D memories present obstacles that must be overcome to use, but they provide significant opportunities to help ameliorate memory wall challenges.
- Continued evolution of the scale and latency-sensitive industry-standard or standard-compatible/interoperable interconnects (for example, scale-up merging with scale-out) will occur.
- Increasingly integrated photonics as a means of power reduction, packaging simplicity, and bandwidth enhancement will be seen.
- Improvements in reliability are driven by the need to address resilience (or fault tolerance) at all levels of the system, from hardware to system software to applications.

These technology changes will result in a new macro-political landscape that may influence decisions on next-generation supercomputer procurement. For example

- AI will drive technology directions/priorities, including reduced precision, systolics, and fixed function units.
- Silicon transistor devices are approaching hard limits in scaling, with limited improvements in performance through silicon CMOS scaling, which has implications for specialization, tight integration, and power reduction. These limits introduce a need for deeper co-design alongside other major market forces, such as AI.
- New computing technologies are being explored, including quantum, neuromorphic, and other accelerators that may substantially change the landscape in terms of scaling, reliability, power, and cooling.
- Research in new nonvolatile memories (NVMs) has been occurring for many years. If that work leads to successful productization, it may affect the way we design storage, conduct checkpointing, and in general, manage memory.
- New algorithms (potentially AI inspired and enabled by new accelerators) can also impact performance and scale.

ARCHITECTURE

Two main architectural changes have brought AI and HPC applications closer together. The first is the addition of high-performance GPUs alongside high-performance CPUs for compute, and the second is AI's need for fast and efficient communication within and between compute elements.

One of the biggest shifts in the last decade has been the widespread adoption of GPUs for computation. While accelerated by AI use cases on supercomputers, this trend was occurring independently on HPC systems due to the need for higher compute capabilities while keeping power manageable. Similar motivators (that is, raw performance, performance per watt, performance per area, and performance per dollar) will likely drive the inclusion of accelerator technology (for example, Cerebras, NextSilicon, SambaNova, and potentially quantum or neuromorphic), though the intercept of the latter two's productive use will require additional time.

As GPUs became dominant, the primary architecture of the system remained homogeneous by node. That is, while each node was heterogeneous (micro-heterogeneity), the overall system was homogeneous. Many of these new accelerators are not as general purpose as GPUs, and therefore, systems are likely to be macro-heterogeneous. What remains open is the tightness of coupling of these macro-heterogeneous partitions.

The severity of the memory bottleneck in generative AI has led to other forms of acceleration reentering consideration, including computation near memory (CNM) as well as processing in memory (PIM). These computational accelerators, coupled with collective acceleration in the network, data processing units (DPUs), and forms of compute near storage, create a more diverse acceleration landscape than that enabled by GPUs. Further, as chiplet-based design points lead to finer-grained customization, the opportunity to intermingle compute acceleration with general purpose compute may become attractive to better balance system performance, power delivery, and thermal dissipation.

A major block to heterogeneity, whether it be at the micro or macro level, is the programming model. Without a productive programming model that enables efficient offload to accelerators, the additional hardware will not provide a good return on area, cost, or power investment. The transition from CPUs to GPUs was made easier via a programming model and tool stack for GPUs, and any accelerators will have to match those capabilities to be viable. For example, circuits for CNM have been known for more than 50 years,[14] but the general programmability problem remains unsolved and generally avoided as "too hard" to solve.

One of the significant challenges in the post-exascale era is communication. This challenge involves moving data from memory to compute and between compute. One way to help address this challenge is to move to more tightly coupled architectures. Memory stacking, 2.5D or 3D, has the potential to reduce power and increase bandwidth between compute and memory.

THE FUTURE WILL DETERMINE IF ESS LEADS TO A COMMON STACK ACROSS THE COMMUNITY OR SPLINTERS THE COMMUNITY.

An important aspect of heterogeneous node architectures is moving data between compute elements, specifically between the main CPU and the accelerator. Coarser parallelism leads to less frequent data movement and more efficient use of the accelerator. Traditional HPC applications need serial cores, and many large AI applications are also increasingly benefiting from the utilization of CPUs. Further, many HPC applications remain bulk synchronous with branchy and data-dependent code between parallelizable kernels; that code runs better on CPUs. The AMD Instinct MI300A accelerated processing unit (APU) brings the CPU and the accelerator computing elements together both physically, via chiplets, and programmatically through a unified memory model; Nvidia's Grace-Hopper provides similar benefits using a full reticle CPU and GPU interconnected through NVLink—a chip-to-chip technology. However, hardware and software challenges, such as software offload launch latencies, remain. Tighter coupling may further help improve performance. For example, 3D stacking would allow more memory bandwidth than 2/2.5D integration.

Moving across compute within the same package or same node offers challenges, but significant performance cliffs occur when moving from high-performance nodes to the network due to lower network byte/flop ratios, high network latencies, and high costs of synchronization across nodes. These inefficiencies require application developers to partition their codes in a coarse-grained manner into serial and parallel compute phases, memory movement phases, and network communicator phases with each one optimized independently. This requirement not only impacts programmer productivity but misses opportunities to optimize power efficiency and memory access across the system. These network inefficiencies also limit strong scaling. The bandwidth and latency cliffs are not the only inhibitors of performance. The model of how memory is accessed can also have a large and potentially greater impact on the performance of applications when they communicate outside the node. The right internode memory model with enhanced capabilities, such as atomics and load/store access to memory within a supernode, pod, or hypernode (collections of tightly coupled nodes with an enhanced memory model), can improve strong-scaled performance by more than an order of magnitude. Nvidia's NVLink and the UALink standard (which AMD is a part of) are specific solutions that can provide tighter coupling between nodes. The general UALink industry-standard effort is moving to create an interoperable fabric for these needs. Competing pressures on interconnects will likely move future interconnects from low radix high diameter to high radix low diameter to improve efficiencies across a wide spectrum of use cases.

Two decades ago, the connection model was flat. A core comprised a node, and each node had a network connection. The topology varied (for example, butterfly, hypercube, or torus), but all compute elements were uniformly separated. With the introduction of multiple cores per chip, multiple chips within a node, and multiple GPUs within a node, two levels of connectivity, inter- and intranode, were introduced. This architecture provided a communication latency and

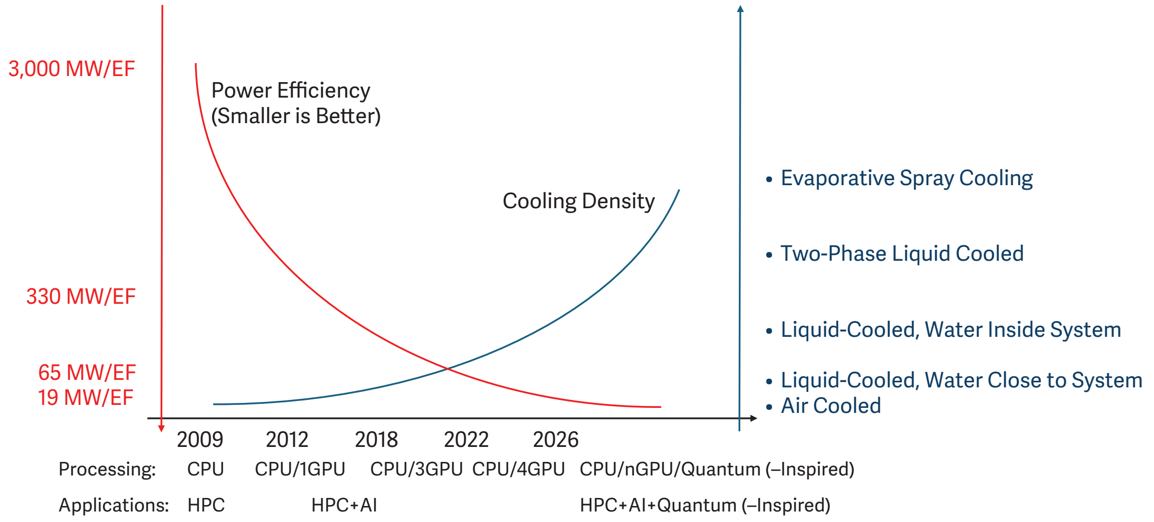


FIGURE 1. Supercomputer power efficiency and cooling over the years.

bandwidth advantage between these computing elements that were contained within a node. However, the architecture came at a cost. Applications—and particularly communication runtimes—needed to be aware of the topological structure to exploit it.

Motivated by AI, scale-up networking is creating another layer in the communication hierarchy. Pods, super nodes, wafer scale, or hypernodes, represent an opportunity to connect tens to hundreds (perhaps small thousands) of nodes in a more tightly coupled manner with memory semantics (for example, load/store access and atomic operations). These architectures have better performance for AI and strong-scaled applications but also introduce a programmability cost. Again, the software layers have an opportunity and responsibility to attune the application appropriately for the communication hierarchy.

An open question remains as to the best overall system architecture since this intermediate communication layer (that is, scale-up: between or within a node and across the whole machine) is more expensive from a cost and power perspective than a flat communication architecture. One possibility that shows promise is merging the connectivity emanating from a node into either scale-up or scale-out connectivity. While this approach is a promising notion, no obvious technologies enable it, yet, but the two main standards initiatives in this space, UALink and the Ultra-Ethernet Consortium (UEC), are currently working on it.

Traditionally, the HPC community relied on large-scale hard-drive-based parallel file systems, such as Lustre and GPFS. In recent times, object store file systems optimized for NVM technology, such as DAOS, VAST, and Weka, are gaining popularity and will increase, including the model stores for AI, such as vector databases. Cloud services have innovated object interfaces, such as S3, that AI frameworks use natively.

FACILITIES

Energy has been driving exascale supercomputing as one of the primary constraints. From the beginning of exascale planning, the desire to keep the spending on power to a minimum led to a target of 20 MW.[15,16] This impacted system designs, specifically cooling, space (the number of racks), and the CPU/GPU ratio. Air cooling was not sufficient, and liquid cooling has become the standard solution for capability-class supercomputers and is seeing broader-based adoption.

Figure 1 notionally presents the evolution of power efficiency on the left (red curve) versus cooling choices on the right (blue curve) during the past few decades. Power efficiency numbers were taken from Oak Ridge National Laboratory supercomputers (Jaguar, Titan, Summit, Frontier). Due to 3D chips, the power density will continue to increase (more than double from 2021 to 2031) according to the IRDS Roadmap,[17] which will require further innovation in cooling, such as immersive or evaporative spray cooling.

In the longer term, both cooling and power requirements may change substantially. Multiple reasons led to the 20-MW limit in the requirements for exascale supercomputers, including cost and the ability to deliver that much power. The new means of energy production, such as small modular reactors (SMRs), are competitively priced per MW and complemented by onsite renewable energy production (for example, wind and solar). If they succeed, they will address both the cost and power delivery to data centers.[18] The AI compute demand and the boom have further shifted the economics and scale of power generation, altering availability and pricing.

SOFTWARE STACK

The system software stack, as defined by everything below an application and above the hardware, continues to increase in complexity. From a modeling and simulation perspective, as the desired capability has increased, system implementers have increasingly turned toward leveraging open source to provide this capability. This change complicates comprehensive testing. The combinations of open source components exponentially increase the number of possible permutations of the software stack. Insufficient connectivity between these open communities (and interest in being connected) has made comprehensive validation significantly more challenging than when a vendor owned all, or most of, the components in a stack.

OpenHPC created a complete and comprehensive system general software stack. Extreme-scale Scientific Software Stack (E4S) of the Exascale Computing Project (ECP) made strides toward unifying the development environment across many open source components. The High Performance Software Foundation (HPSF), unified by Spack, is making strides toward providing optimized software stacks for well-defined systems. Nonetheless, challenges remain, and a stronger community testing effort, perhaps under HPSF, is still needed.

The inclusion of AI software stacks on supercomputers has significantly increased the number of components of the overall software stack. More importantly, AI infrastructure, including the software stack, is undergoing rapid change. The key contributors are investing significant effort to support this rapidly evolving environment while other organizations are challenged to keep up. Overall, the rapid evolution limits the organizations that can stand up and maintain an AI stack, which further increases the need for community efforts toward testing and maintaining the overall software stack.

While E4S was United States centered, Europe is developing the European Software Stack (ESS). The EuroHPC JU will work with stakeholders to coordinate co-design in the research and investigation of hardware and software activities and ensure that those activities meet user requirements and that developed technologies are deployed. Funding is planned for the different building blocks in HPC, AI, and quantum computing (QC) from innovation to deployment, targeting different technical readiness levels as required by the status of hardware developments. Europe will focus on multiple aspects, such as performance and efficiency, AI-software integration, energy consumption, workflow managers, and support to European processors, among others. The future will determine if ESS leads to a common stack across the community or splinters the community.

As discussed previously, macro-heterogeneity is on the horizon; enhancements of the software will be needed to incorporate the new elements into the system as well as to support macro-heterogeneity generally. To make these accelerators productive, a comprehensive software stack will need to be developed to enable nonexpert application developers. User interfaces, libraries, debuggers, validation tools, high-level programming models, and languages are needed as well as compilers to translate high-level languages to be distributed over coarse-grain reconfigurable architectures or to QC circuits and transpilers that adapt already-compiled circuits to a dedicated technology.

As the software stack becomes more complex and the overall user code moves from a single executable to a complex set of interconnected executables, we will need an overarching workflow infrastructure. Some examples of workflow management exist today, but those capabilities will need to be enhanced to cover the great variety of emerging software stacks. They will also require many new capabilities, such as the control of data movement and enhanced authentication, security, and monitoring.

The amount of power consumed by supercomputers is reaching an inflection point where the cost of electricity throughout the life of the system is approaching its capital cost. New software capabilities must be created to enable users to understand and optimize the tradeoff between performance and energy (for example, to allow a user or system administrator to reduce performance by 10% to save 40% on energy). We will also need support to ramp up and down power more smoothly to meet the requirements of electricity providers.

OPERATIONS

The U.S. ECP was a multibillion-dollar effort, with multiple hundred-million-dollar procurements. In addition, the cost to operate an exascale supercomputer is on the order of 100 million U.S. dollars, a significant part of its total cost of ownership.

Producing and procuring a capability-class supercomputer is a complex operation that is not optimal for the participants in the procurement: regulators, users, integrators, and suppliers. Distributed spending with incremental upgrades could be beneficial. Similarly, the operating expense costs are becoming too high to be financially sustainable. New means of producing and delivering supercomputers could prove beneficial for multiple parties.

Current supercomputers are designed to run applications at an extreme scale. While needed for capability-class applications, this model has challenges for maintenance and partial system refreshes. Accelerator road maps are also more frequent and shorter than the lifetime of supercomputers, which makes refreshes more desirable than in the past, from both the performance and power/cost perspective.

NONFUNCTIONAL REQUIREMENTS

Reliability has long been a focus of traditional HPC, extending from high-level software to ensure that it did not have any single points of failure, down to the silicon, including both compute and memory. This focus was needed as the high-level fault tolerance model in applications was that if one node failed, the entire application failed. Thus, as the machine grew in node count, it was imperative that reliability was improved. Nevertheless, the mean time between failure on the largest supercomputers has dropped from around a week on emergent petascale systems to a handful of hours on emergent exascale systems. With each generation, new points of hardware and software reliability failures emerge due to ever increasing hardware complexity and software not planning for significant implications of heterogeneous architecture implementations.

Innovations in checkpointing architecture in conjunction with improved bandwidth for checkpoints have predominantly ameliorated the impact that this decreased reliability has on system availability. However, unless something changes, this trend will be unsustainable for the next three orders of magnitude of system performance improvement. Fewer applications can productively employ a full exaflop of compute than the number that could employ a full petaflop. This potentially implies a different usage model for supercomputers in the next decade. Each facility's workload will determine whether petascale or exascale resources (for example, compute, memory capacity, and memory bandwidth) are needed.

AI has only recently been run at large scales. Thus, GPUs have not focused as much on reliability as CPUs that were designed for supercomputers. The AI software stack has also not had years of focus on reliability and ensuring no single point of failure. Recent data from Meta,[19] Alibaba, Google,[20] and others show the consequences. As AI continues to scale and systems become larger with the desire to run capability-class applications, an increased focus on fault tolerance will be needed, both in designing and implementing more reliable hardware and in changing the application fault tolerance model.

AI applications are inherently more resilient to failures because of the nature of their computation. While academic work has explored application-level fault tolerance for modsim applications, it has not been implemented in practice as most of the work could address only specific computational kernels rather than the resilience of the entire application. In one form or another, reliability will need more focus moving forward.

SUMMARY AND OUTLOOK

In this article, we presented our predictions of the future of supercomputing. We first discussed increased use and adoption, followed by evolving technologies and workloads. We then presented the architecture, facilities, software stack, operation, and nonfunctional

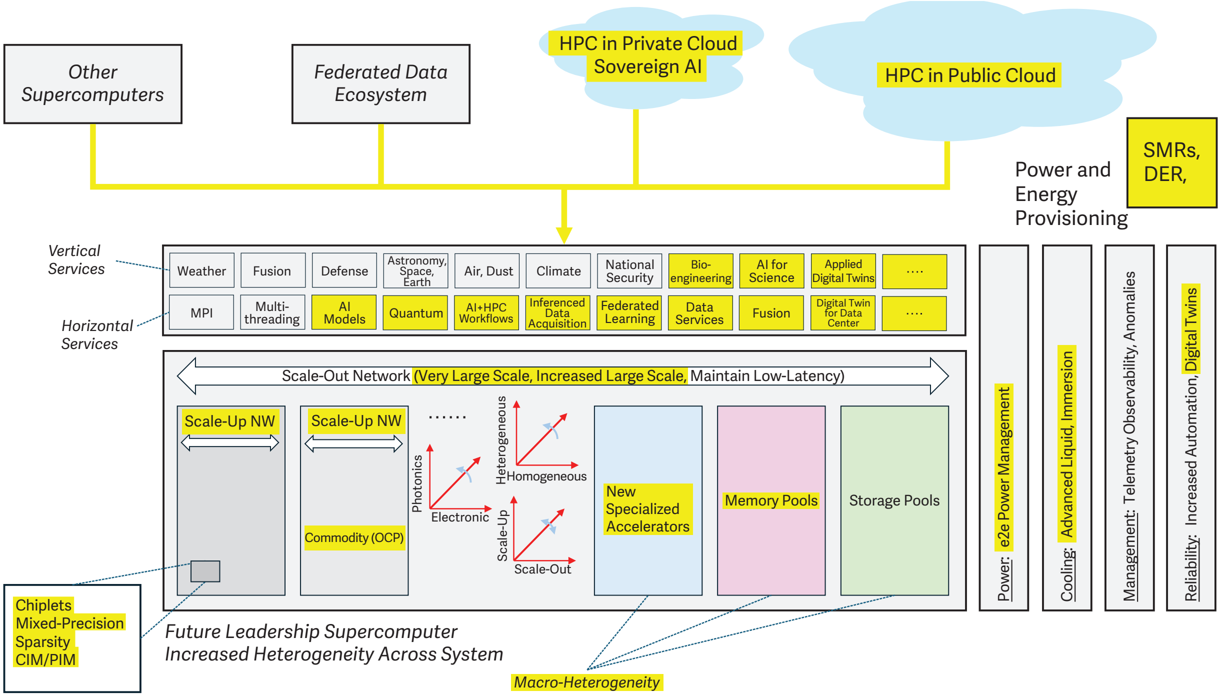


FIGURE 2. High-level supercomputing architecture. Highlighted text in yellow represents new features compared to existing supercomputers. API: application programming interface; SMRs: small modular reactors; DER: distributed energy resources; CIM: computing in memory; OCP: open compute; e2e: end-to-end.

requirements. We concluded with some recommendations to critical actors in supercomputing.

Figure 2 and Table 1 summarize our predictions. Figure 2 describes the architecture of future supercomputing, emphasizing the innovations required. Table 1 succinctly presents the evolution of HPC over decades, from traditional to future supercomputing.

Achieving the next level of scale will require innovation, just like it did to get from petascale to exascale. This innovation will likely need to come across the whole system, including new accelerators, interconnects, system software, application and algorithmic innovations, and power and cooling. Some of the scaling may be possible to achieve by leveraging macro-heterogeneity, for example, through the use of AI-specific, quantum or quantum-inspired, or other accelerators in the context of a more traditional GPU-based supercomputer.

Supercomputers will also benefit from the growth in the bandwidth of interconnects. Photonics could help overcome limited processor shoreline performance, power, and packaging. However, additional investments will have to be made to avoid congestion at scale and to address both jitter and tail latency.

In terms of power and cooling, the current limitations will remain and will have to be addressed with onsite power generation, possibly with SMRs and renewable energy sources as complements to grid supplies. Cooling will require new techniques, as discussed in the "Facilities" section. Locating data centers in zones where power is cheap and reliable can also help. Areas with abundant water and favorable climates will assist with cooling challenges.

Sustainability is challenging in supercomputing due to the extreme use of power. Some of the approaches of large-scale enterprise data centers can be applied (for example, following the sun or server consolidation) to a limited extent. Sustainability awareness can help, as can using digital twin techniques to conduct what-if-analyses and understand where opportunities lie.

The use of AI is inherently tied to ethics and is an important topic that will need to be addressed given the widespread use of AI. AI is effective at improving productivity in software development. Productivity in

TABLE 1. Comparing approaches to building and consuming leadership supercomputing systems.

Comparison criteria		Supercomputer eras						
		Traditional HPC supercomputer (1990 to present)*	Grid[21] (2000–2010)	Cloud[22] (2006 to present)	AI cluster training	AI cluster-inference	HPC cloud[23]	Future HPC supercomputer
How the system is built	Coupling	Very Tight scale-out	Tight scale-out federated	Loose	Tight (scale-up and scale-out)	Loose + scale-up	Loose, scale-up, medium-tight	Configurable
	Scale	<10× exascale	Multisite (federated?)	Multiregions	Collocated	Distributed	Multiregions	>100× Exascale
	Reliability	Job-based restarts	Job-based restarts	Cloud like†	Job-based restarts	Cloud like†	Cloud like	Job restarts + cloud like*
	Elasticity	No	Desired	By design	Moderate	Cloud like	Cloud like	Generally desired, essential for broader workflows
	Storage System	Parallel FS (write intensive)	Grid FS	Block, object (read intensive)	Read-training data write-ckpt (file, object)	Read intensive (mostly objects)	Block, object (read intensive)	Mixture
Consumption model	Business Adoption	Governments	Governments/industry	Consumer/enterprise	Model builders, sovereign AI	Service providers, enterprise	Government/industry/provider	Converged AI + HPC users
	Networking‡	No (@ Periphery§)	Yes	Inherent	Yes	Yes	Yes	Yes
	Multitenancy	Minimal	Yes	Inherent	Moderate (job based)	Yes	Inherent	Yes
	Virtualization	No (well, some containers)	Some	Built-in VMs	Containers + K8s	Containers + K8s	Built-in VMs, containers	Yes
	Optimized for	Mod/sim HPC	(data-intensive) HPC	Content serving, horizontal scale	Training and tuning large AI models	Models at scale, agents, workflows	Loosely coupled HPC, AI	HPC, AI

VMs: virtual machines.

* While the first supercomputer was delivered in 1964, we started counting from 1990 when the first modern scale-out computer was delivered.

† Cloud-like reliability: 1) stateless/fungible VMs; 2) reliable persistence layer (S3, etc.); 3) restartable service requests; and 4) eventual consistency for distributed tasks.

‡ Most HPC and AI training is dominantly East-West, while cloud and AI serving are dominantly North-South (N-S). The difference with AI is that it is N-S + scale-up (multi-GPU networks), while the traditional cloud is largely N-S.

§ Supercomputers are connected to the outside—but only at the periphery of the system, with a different network.

developing supercomputing applications is critical but also hard to automate using AI due to the performance and scale requirements.

RECOMMENDATIONS

We make recommendations to key actors in the supercomputing ecosystem: supercomputer centers, developers, scientists/users, and industry.

Our recommendations for supercomputer centers are as follows:

- Workloads of the future will continue to have demands for tightly coupled, highly parallel, and noise-free infrastructure at scale. Therefore, the growth in the needed capabilities of future supercomputers will continue, and centers should continue to plan to procure them.
- Future supercomputers may be supplemented by leveraging offload to a public or private cloud or large AI infrastructures for training or services that enhance productivity. Centers

should investigate how to incorporate complex workflow capabilities that allow this interaction as well as intrafacility and interfacility workflows. Infrequent delivery of single large supercomputers puts a strain on providers, users, and maintainers of supercomputers. An alternative incremental delivery should be explored to ensure smooth delivery and secure a more reliable introduction of new features. It also puts HPC at a disadvantage from a performance standpoint. GPU performance is still scaling rapidly, and AI is forcing an acceleration in hardware innovation from compute to networks.

Our recommendations for developers and the open source community are as follows:

- Most of the system software running on supercomputers is becoming open source. The community should become more strategic about planning and delivering new features and secure approaches and infrastructures to be able to develop and test solutions at scale.
- To allow the broadest productive use of software, instilling good software engineering practices into community code will be beneficial (for example, the work E4S did made its components more accessible to a wider community). HPSF is a good step in this direction.
- As AI is becoming more prevalent in almost every aspect of programming, the models should be treated the same way as open software. The data that were used for training should be made available and documented. While enhancement based on private data will be necessary for some use cases, the data on which open models are based must also be open.
- In general, but especially for science applications, focus on the explainability of AI methods.
- Open hardware is becoming an alternative that needs to be carefully evaluated and considered in supercomputing solutions. Open firmware is also an interesting direction to enhance security and maintainability.
- Work on leveraging low-precision hardware to emulate or perform high-precision calculations is essential. Ultimately, scientific applications need a more rigorous error-based approach to numerical precision.

Our recommendations for scientists and users of supercomputers are as follows:

- Adjust to using cloud infrastructure and AI programming models combined with the existing traditional HPC algorithms.
- Continue to be innovative in terms of continuously increased scale and alternative programming models offered by new hardware (for example, AI accelerators and quantum).
- Invent new algorithms and applications to leverage the new AI and future computing and memory technology.

Our recommendations for industry, integrators, and system vendors are as follows:

- Ensure sufficient interoperability across the components and interconnects to enable reusability across supercomputers.
- Provide sufficient documentation and interfaces for using hardware and core system software.
- Support interfaces and software for the maintenance and management of supercomputers at scale.
- Provide the capability to combine AI capability productively into existing applications.

The need for supercomputing continues to grow. In addition to the needs of traditional scientific computing, AI's needs are driving the evolution of computing hardware and software. The authors lay out several challenges and opportunities for the next decade for computing facilities; developers, scientists and users; and industry.

ACKNOWLEDGMENT

We would like to thank Cullen Bash, Paolo Faraboschi, and Samantika Sury for their advice and insights that substantially helped improve this article. This manuscript has been authored in part by Lawrence Livermore National Security, LLC under Contract DE-AC52-07NA2 7344 with the U.S. Department of

Energy (DOE) (Grant LLNL-JRNL-2005193); it is based in part upon work supported by the DOE, National Nuclear Security Administration under Contract 89233218CNA000001; and it used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725. LA-UR-25-23959 (approved for public release). Rosa M. Badia acknowledges funding from Projects CEX2021-001148-S, and PID2023-147979NB-C21 from the MCIN/AEI and MICIU/AEI/10.13039/501100011033 and by FEDER, UE, by the Departament de Recerca i Universitats de la Generalitat de Catalunya, research group MPiEDist (2021 SGR 00412). AMD, the AMD Arrow logo, AMD CDNA, AMD Instinct, AMD ROCm, AMD Infinity Cache, AMD Infinity Fabric, and combinations thereof are trademarks of Advanced Micro Devices, Inc. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies.

REFERENCES

1. R. M. Badia, I. Foster, and D. Milojicic, "Future of HPC," *IEEE Internet Comput.*, vol. 27, no. 1, pp. 5–6, Jan./Feb. 2023, doi: 10.1109/MIC.2022.3228323.
2. D. Milojicic, P. Faraboschi, N. Dube, and D. Roweth, "Future of HPC: Diversifying heterogeneity," in *Proc. Design, Automat. Test Europe Conf. Exhib. (DATE)*, 2021, pp. 276–281, doi: 10.23919/DATE51398.2021.9474063.
3. N. Dube, D. Roweth, P. Faraboschi, and D. Milojicic, "Future of HPC: The internet of workflows," *IEEE Internet Comput.*, vol. 25, no. 5, pp. 26–34, Sep./Oct. 2021, doi: 10.1109/MIC.2021.3103236.
4. G. M. Shipman et al., "The future of HPC in nuclear security," *IEEE Internet Comput.*, vol. 27, no. 1, pp. 16–23, Jan./Feb. 2023, doi: 10.1109/MIC.2022.3229037.
5. E. Deelman et al., "High-performance computing at a crossroads," *Science*, vol. 387, no. 6736, pp. 829–831, 2025, doi: 10.1126/science.adu0801.
6. R. Stevens, V. Taylor, J. Nichols, A. B. Maccabe, K. Yellick, and D. Brown, "AI for science: Report on the Department of Energy (DOE) Town Halls on artificial intelligence (AI) for science," Argonne National Lab. (ANL), Argonne, IL, USA, Tech. Rep. ANL-20/17; 158802; TRN: US2103893, Feb. 2020.
7. W. Jia et al., "Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning," in *Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal.*, Atlanta, GA, USA, 2020, pp. 1–14.
8. S. Das et al., "Large-scale materials modeling at quantum accuracy: Ab initio simulations of quasicrystals and interacting extended defects in metallic alloys," in *Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal.*, Denver, CO, USA, 2023, pp. 1–12.
9. G. Dharuman et al., "MProt-DPO: Breaking the ExaFLOPS barrier for multimodal protein design workflows with direct preference optimization," in *Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal.*, Atlanta, GA, USA, 2024, pp. 1–13.
10. H. Ootomo, K. Ozaki, and R. Yokota, "DGEMM on integer matrix multiplication unit," 2024, *arXiv:2306.11975*.
11. J. Athavale et al., "Digital twins for data centers," *Computer*, vol. 57, no. 10, pp. 151–158, Oct. 2024, doi: 10.1109/MC.2024.3436945.
12. W. Brewer et al., "A digital twin framework for liquid-cooled supercomputers as demonstrated at exascale," in *Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal.*, Atlanta, GA, USA, 2024, pp. 1–18.
13. L. R. Leung, "Earth system modeling or actionable science," Pacific Northwest Nat. Lab., Richland, WA, USA, Jul. 2024. [Online]. Available: https://www.nersc.gov/assets/Uploads/NERSC_Leung_2024_final.pdf
14. R. R. Seeber, "Associative self-sorting memory," *presented at the Eastern Joint IRE-AIEE-ACM Comput. Conf. (IRE-AIEE-ACM)*, New York, NY, USA: ACM, Dec. 13–15, 1960, pp. 179–187.
15. P. Kogge et al., "ExaScale computing study: Technology challenges in achieving exascale systems," DARPA, Arlington, VI, USA, Sep. 2008. [Online]. Available: https://ftp.eecs.berkeley.edu/~yelick/papers/Exascale_final_report.pdf
16. S. Atchley et al., "Frontier: Exploring exascale the system architecture of the first exascale supercomputer," in *Proc. Int. Conf. High Perform. Comput., Netw., Storage Anal.*, Denver, CO, USA, 2023, pp. 1–16, doi: 10.1145/3581784.3607089.
17. "International Roadmap for Devices and Systems™ 2023 update: Systems and architectures," IEEE, Piscataway, NJ, USA, 2023. [Online]. Available: https://irds.ieee.org/images/files/pdf/2023/2023IRDS_Perspectives.pdf
18. C. Bash, J. Bian, D. Milojicic, C. D. Patel, L. Strezoski, and V. Terzija, "Energy supplies for future data centers," *Computer*, vol. 57, no. 7, pp. 126–134, Jul. 2024, doi: 10.1109/MC.2024.3393248.

19. A. Grattafiori et al., "The Llama 3 herd of models," 2024, *arXiv:2407.21783*.

20. N. Jouppi et al., "TPU v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings," in *Proc. 50th Annu. Int. Symp. Comput. Archit. (ISCA)*, New York, NY, USA: ACM, 2023, pp. 1–14.

21. C. Kesselman and I. Foster, *The Grid: Blueprint for a New Computing Infrastructure*. Burlington, MA, USA: Morgan Kaufmann, 1999.

22. A. Gupta et al., "Evaluating and improving the performance and scheduling of HPC applications in cloud," *IEEE Trans. Cloud Comput.*, vol. 4, no. 3, pp. 307–321, Jul./Sep. 2016, doi: 10.1109/TCC.2014.2339858.

23. T. Gamblin et al., "HPC center of the future: R&D acquisition intent," Lawrence Livermore Nat. Lab., Livermore, CA, USA, Tech. Rep. LLNL-TR-871269, Nov. 14, 2014.

SCOTT ATCHLEY is the CTO at Oak Ridge National Laboratory's National Center for Computational Science, Oak Ridge, TN 37830 USA. Contact him at scott@ornl.gov.

ROSA M. BADIA is a workflow and distributed computing manager at the Barcelona Supercomputing Center, 08034 Barcellona, Spain. Contact her at rosa.m.badia@bsc.es.

BRONIS R. DE SUPINSKI is the CTO for Livermore Computing at the Lawrence Livermore National Laboratory, Livermore, CA 94550 USA. Contact him at bronis@llnl.gov.

JOSHUA FRYMAN is a fellow at Intel, Hillsboro, OR 97124 USA. Contact him at joshua.b.fryman@intel.com.

DIETER KRANZLMÜLLER is the chair of the board of directors at the Leibniz Supercomputing Centre (LRZ), 85748 Garching bei München, Germany. Contact him at dieter.kranzlmueller@lrz.de.

SRILATHA MANNE is a senior fellow at Advanced Micro Devices, Inc., Seattle, WA 98103 USA. Contact her at srilatha.manne@amd.com.

PEKKA MANNINEN is the director of science and technology at CSC, the Finnish IT Center for Science, 02101 Espoo, Finland. Contact him at pekka.manninen@csc.fi.

SATOSHI MATSUOKA is the director of THE RIKEN Center for Computational Science, Saitama 351-01, Japan. Contact him at matsu@acm.org.

DEJAN MILOJICIC is an HPE fellow and vice president at Hewlett Packard Labs, Milpitas, CA 95035 USA. Contact him at dejan.milojicic@hpe.com.

GALEN SHIPMAN is a computer scientist at the Los Alamos National Laboratory, Los Alamos, NM 87545 USA. Contact him at gshipman@lanl.gov.

ERIC VAN HENSBERGEN is a fellow at ARM, Austin, TX 78735 USA. Contact him at eric.vanhensbergen@arm.com.

ROBERT W. WISNIEWSKI is an HPE fellow, chief architect, and vice president of AI and HPC Solutions at Hewlett Packard Enterprise, Spring TX 77389 USA. Contact him at robert.wisniewski@hpe.com.

EDITOR: Tom Coughlin, Coughlin Associates; tom@tomcoughlin.com

DEPARTMENT: MEMORY AND STORAGE

This article originally appeared in Computer vol. 57, no. 3, 2024

How Emerging Memories Extend Battery Life

Jim Handy, *Objective Analysis*

Tom Coughlin, *Coughlin Associates, Inc.*

Energy consumption is an issue with many connected digital products. Resolving energy efficiency issues and putting more memory in less die space create opportunities to use new nonvolatile memories for code storage and cache memory.

With today's explosion of battery-operated devices like industrial and consumer Internet of Things (IoT) endpoints, wearables, health monitors, and such, a growing focus is on the energy consumption of these devices.

INTRODUCTION

Designers must weigh tradeoffs among functionality, portability, and battery life since each of these play off against one another. For example, a device can have very elaborate functionality and a long battery life if a large battery is used, but that makes the device less portable. With a smaller battery the battery life will be shortened, but the product becomes more portable. If the designer strips down the feature set, then the smaller battery might do the job for a reasonable time.

Interestingly, this tradeoff is now being helped out through the use of new nonvolatile memory types, which have only become widely available in the past few years: magnetic RAM (MRAM), resistive RAM (ReRAM), phase-change memory (PCM), and ferroelectric RAM (FRAM). This article will examine those technologies and will show how their use can optimize the balance of these tradeoffs.

CURRENT MODEL

For the past few decades, endpoints have tended to use the same memory types: NOR flash, SRAM, and, in some cases, dynamic RAM (DRAM), to support the central processor. Often the NOR flash and SRAM are integrated into the processor chip in the form of a microcontroller unit (MCU). Some systems increase the density of these memories by adding external discrete NOR and SRAM chips, which adds to the cost. If an external NOR is used, some of its contents will often be stored within the MCU in an SRAM cache since NOR reads are relatively slow compared with the speed of a program's execution. Furthermore, SRAM scales more slowly than CMOS logic, and that means that the relative cost of the cache increases over time to become a growing share of the MCU chip's cost.

This works well in applications where there is no need to store data, which is done either to recover cleanly from power interruptions or to allow the chip to be powered down for energy saving. Things become more difficult when data must be stored. The designer usually chooses between two options:

- Use a battery-backup SRAM.
- Write data into the NOR flash.

A battery-backed SRAM works very nicely, as long as the battery functions. During normal operation, the SRAM operates at full speed, and it consumes very little power when in standby mode. Unfortunately, batteries have a limited lifetime and so must be changed. If the device needs to maintain the SRAM's data through this battery change, then the design becomes much more elaborate. This approach is easy until maintenance is considered, and then the design becomes significantly harder.

Digital Object Identifier 10.1109/MC.2023.3340799
Date of current version: 1 March 2024

Some MCUs include a battery-backed SRAM, and this can simplify the designer's task a bit. Still, the battery replacement issue becomes a challenge. Also, as mentioned above, SRAM is not scaling with CMOS logic, so the SRAM's cost will become an increasing part of the MCU's cost over time.

NOR flash does not need a battery to store information, making it more appealing thanks to its lower complexity, but NOR flash takes significantly more time and energy to perform a write and has to go through a block erase if space does not yet exist for that write. A memory address cannot simply be overwritten in flash.

For example, while a NOR flash chip might take a certain amount of energy to read a page, the page programming might take 15 times as much energy due to higher voltages and currents along with longer cycle times. However, that is only true if there is free space for the data to be written into. If a block must be erased to provide room for that write, then the whole erase-then-write process can consume about 20,000 times as much energy as a read.

Furthermore, embedded NOR flash stops scaling at 28 nm. The advent of the fin-shaped field-effect transistors processes at 14 nm gets in the way of producing NOR flash, so foundries that produce aggressive process geometries either are in development or have already developed other nonvolatile memory technologies to replace NOR at 14 nm and smaller process nodes.

EMERGING MEMORIES AS A SOLUTION

Those new memory types that were mentioned at the beginning of the article, MRAM, ReRAM, PCM, and FRAM, all have attributes that make them better than either battery-backed SRAM or NOR flash for data storage and are poised to become a lower-cost alternative to either SRAM or NOR flash. All offer very fast read and write, all promise to scale to process nodes beyond those supported by NOR flash and SRAM, and all can help the engineer design a lower-energy system than SRAM or NOR.

One benefit that has not been mentioned so far is the ability to power a system down at any time without needing to move data from volatile RAM (SRAM or DRAM) into a nonvolatile memory. While systems with battery-backed SRAM can simply leave the SRAM in a standby state, running off the backup battery's power, other systems must move data from RAM into NOR flash, and this consumes a lot of energy. With an emerging memory technology, the same architecture can be used as with a battery-backed SRAM: the data can remain where they are at power-down to be accessed again when power is restored. This lends itself to a power-saving approach that Intel calls "Hurry Up, Get Idle" ("HUGI"). The system is designed to find opportunities to shut down frequently, saving valuable battery energy when it is powered down.

THE SYSTEM IS DESIGNED TO FIND OPPORTUNITIES TO SHUT DOWN FREQUENTLY, SAVING VALUABLE BATTERY ENERGY WHEN IT IS POWERED DOWN.

The drawback is that none of these technologies is yet produced in the kind of volume that will drive their costs down. Current memory technologies, like DRAM and NAND flash, are produced in high enough volumes and have been produced for so very long that manufacturers understand how to drive the costs out of the production process. This is not the case with newer memory technologies, so today they are the higher-cost alternatives. From the perspective of production volume, these technologies are still very young, even if they may have been in production for a number of years.

Fortunately, the migration to sub-28-nm process technologies is increasing the production volume of these memory types, which will eventually lead to cost reductions. In the end, this promises to make these technologies cheaper than SRAM or NOR flash, but this is not the case today.

While the sub-28-nm problem is unique to NOR flash, SRAM's biggest problem is that each bit is very large since it requires six transistors to implement, while a NOR flash or an emerging memory bit is much smaller, typically taking only a single transistor and, in the case of the emerging memories, some kind of bit storage element (more on those later). In some cases, the bits will be stored differentially to increase speed, but these cells still consist of only two transistors and two storage elements. This makes them necessarily cheaper than SRAM as long as the wafer costs are the same. Today the wafer costs for emerging memories are higher, but that difference will fade as the production volume increases.

EMERGING MEMORY TYPES

The following memory types are in production today.[a] All offer roughly the same attributes, and any one of them could rise above the others to become the leading memory type over the next decade. All of them provide persistence (that is, they are nonvolatile), all write in place (which is a vast improvement upon flash's block erase and page write, with erase before write), all have fast, low-energy writes, and all can scale to smaller process geometries than are currently available. They perform nearly as well as battery-backed SRAM but without the battery and with the promise of becoming much less expensive than SRAM.

MRAM

MRAM comes in several forms. Toggle mode MRAM is in the highest volume today but has trouble scaling past 120 nm, so it is being displaced by spin transfer torque (STT) MRAM. In the future, other versions, mainly spin orbit torque, with faster performance, may replace STT MRAM. Each bit of any of these technologies can be implemented with a single MRAM bit element and a single transistor. Today the transistor's size limits how small a bit can be made since the technology requires relatively high currents, but researchers are working on a solution to this problem.

All MRAM uses a special layer of material that exhibits the property of giant magnetoresistance to store the bit. This material, while produced in high unit volume to manufacture recording heads for HDDs, has a very small die size, so it is not yet manufactured in the high wafer volumes of silicon CMOS and is therefore expensive today.

MRAM is available as a foundry process from TSMC, Samsung, and GlobalFoundries. Discrete MRAM chips are available from Everspin and Avalanche.

ReRAM

ReRAM uses a resistive element to store a bit. While some manufacturers use a less-understood material to produce the bit element, certain companies, namely, Weebit Nano and Crossbar, have developed ReRAM that is based on a slight change to the same silicon dioxide insulation material that is universally used in silicon semiconductors. This should accelerate these technologies' ability to reach the economic benefits of high-volume production.

There are two basic programming mechanisms: filamentary and oxygen depletion. While this article will not explain these mechanisms, neither is as well understood as is standard silicon CMOS.

ReRAM cells consist of a single resistive element and a selector, which today is typically a transistor. This means that the bit size rivals that of MRAM and NOR flash. Future ReRAMs are expected to use a two-terminal selector, which can be built below the resistive element to cut the bit's size in half and which will facilitate layering bits in multiple "decks" to further double, triple, or quadruple the number of bits that can fit into a given area of silicon.

Today discrete ReRAM chips are produced in volume by Fujitsu and its partner Panasonic. Foundries TSMC, Samsung, Global Foundries, Winbond, Skywater, DB HiTek, SMIC, and Crocus Nano all offer an embedded ReRAM process.

PCM

PCM (or PRAM) has had its day in the sun in its 3D XPoint memory incarnation. Like a ReRAM, it stores

[a] Report: Emerging Memories Branch Out, Coughlin Associates and Objective Analysis, 2023. http://Objective-Analysis.com/reports/Emerging#

the bit in a resistive element, but the storage mechanism is different since it involves a material change. In most PCMs, temperature ramps are used to change the storage element between crystalline (conductive) and amorphous (nonconductive) phases, but there is another method that changes the resistance through high programming currents.

PCM is based on chalcogenide glasses, which are not as well understood as is silicon. Some of these glasses also involve elements that are difficult to manage in a silicon fab.

As with ReRAMs, PCM can use either a transistor or a two-terminal selector. The most common two-terminal selector today is also based on a chalcogenide glass, so PCM is a good fit. Intel and Micron were able to use this to their advantage since it allowed multiple "decks" of 3D XPoint memory to be easily stacked, and that reduced the technology's cost for a given memory capacity.

Today, only STMicroelectronics provides PCM as an embedded memory in its "Stellar" microcontroller. BAE sells its PCM "C-RAM" to aerospace applications that value its immunity to radiation.

FRAM

FRAM involves no iron, despite its name. Since this technology stores a bit's state via hysteresis that resembles the ferromagnetic hysteresis loop, researchers called it FRAM. FRAM is also the first semiconductor memory, with the first multibit monolithic prototype developed in 1955, three years before Jack Kilby's 1958 invention of the integrated circuit.[b]

From the 1950s through 2010, all FRAM was produced using either strontium bismuth titanate or lead zirconium titanate, both of which include high-mobility elements that can easily contaminate a silicon fab. This limited their popularity. In 2010, NamLab in Dresden, Germany, found evidence of ferroelectric behavior in hafnium oxide (HfO), which is prevalent as a gate dielectric in very advanced silicon processes; this discovery has led to a lot of research but not yet to any actual products.

Discrete FRAM is produced by Infineon and Lapis Semiconductor, TI embeds it into a microcontroller, and Fujitsu and Panasonic embed FRAM into RFID chips for mass-transit fare cards.

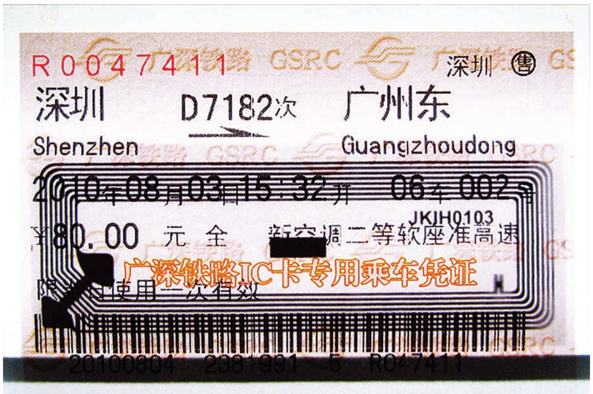


FIGURE 1. A Guangzhudong Shenzen Railway Company fare card. (Source: Wikimedia Commons, IC ticket of Guangshen Railway.jpg.)

LOW-POWER APPLICATIONS OF EMERGING MEMORIES

Here we will present a few of the many applications that use emerging memory technologies to save energy in low-power applications.

Mass-transit fare cards

Very early examples of such applications are the mass-transit fare cards pioneered in Japan and now used broadly in Asia. These cards have no internal power source, yet they store the value assigned to them less any transactions from the card's use. They are read via near-field communications (NFC).

The cards must store the value, allow it to be read, and then allow a new total to be written back into the card, all using only the energy provided by the NFC signal. Fujitsu and Panasonic chose to use FRAM for this application because its fast low-power write could be powered by the NFC signal.

An example of one of these cards is shown in Figure 1. They are the same size and shape of any standard charge card.

Personal fitness monitors

There is widespread use of MRAM in personal fitness monitors, which must perform numerous sophisticated tasks for a full day or more using only the energy that will fit into a small battery within the watch-sized

[b] FRAM Turns 68, The Memory Guy Blog, Jim Handy, 10 July, 2020. https://TheMemoryGuy.com/fram-turns-68/.

FIGURE 2. Google's Fitbit Luxe.

FIGURE 3. Garmin's Versa 4.

device. Many of these use the MRAM version of the Apollo 4 processor from Ambiq, a company that uses subthreshold logic to get the highest performance out of the absolute smallest amount of energy possible.

Figures 2 and 3 show two examples: the Fitbit Luxe from Google (Figure 2) and the Versa 4 from Garmin (Figure 3). Garmin has another device not pictured here, the Fenix 7 Solar, which adds a solar cell to an MRAM-based wearable to further extend the time between charges.

Medical devices and prosthetics

Various development efforts are underway to incorporate emerging memory into everything from disposable health monitoring devices, which look more like a small bandage than instrumentation, up to cardiac defibrillators and hearing aids. While the developers generally do not disclose the chips used inside their devices, we understand that MRAM, ReRAM, and FRAM are all being used in such applications.

BIG CHANGES ON THEIR WAY

Readers should expect to see significant changes leading to longer battery life in the next few years as emerging memory technologies become widespread in IoT endpoints and other battery-operated equipment. There may even be a rise in the use of scavenged power, as is already done in mass-transit fare cards and in Garmin's Fenix 7 solar wearable device.

In the end, a lot of this will be enabled through the use of new memory technologies, which drastically reduce the energy requirements of data storage. These technologies are about to ramp pretty quickly, in support of finer process geometries, so they will become common over the next five years.

JIM HANDY is the general director of Objective Analysis, Los Gatos, CA 95032 USA. Contact him at jim.handy@objective-analysis.com.

TOM COUGHLIN is the president of Coughlin Associates, Inc., San Jose, CA 95124 USA. Contact him at tom@tomcoughlin.com.

EDITOR: Conrad James, Sandia National Laboratories, USA, cdjame@sandia.gov

DEPARTMENT: MICROELECTRONICS

This article originally appeared in Computer vol. 57, no. 4, 2024

Semiconductor Memory Technologies: State-of-the-Art and Future Trends

Shimeng Yu and Tae-Hyeon Kim, *Georgia Institute of Technology*

This article surveys the recent development of semiconductor memory technologies spanning from the mainstream static random-access memory, dynamic random-access memory, and flash memory toward emerging candidates such as resistive, ferroelectric, and magnetic memories. Pathways for future technological innovations are presented.

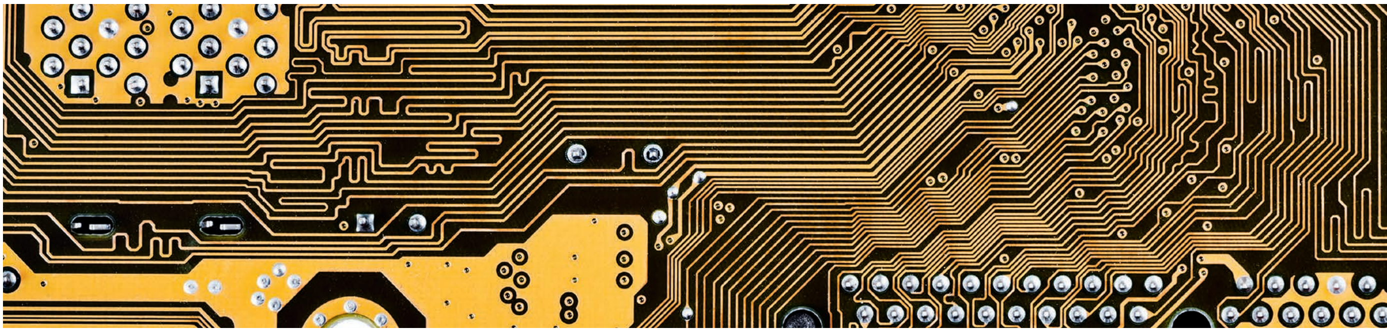
Semiconductor memory technologies play a pivotal role in modern computing systems, serving as the primary means of storing and retrieving digital information. These technologies encompass a diverse range of memory types, each with unique characteristics suited for specific applications. Demand for higher capacity, faster speed, and lower power consumption continues to drive innovation in memory technologies. Additionally, the proliferation of data-intensive applications like artificial intelligence (AI), machine learning (ML), and big data analytics fuels the need for more advanced memory solutions. As a result, the semiconductor memory market is expected to remain robust, with ongoing developments shaping its landscape.

The memory hierarchy traditionally refers to the organization of different types of memory in computing systems, ranging from high-speed, low-capacity registers and caches to slower but larger main memory and persistent storage. However, emerging trends like Compute Express Link (CXL) are blurring the boundaries of this hierarchy. CXL, a high-speed interconnect standard, enables processors to access various types of memory and accelerators as if they were part of the CPU's memory space. This architecture allows for more flexible and efficient data movement between different memory types, including working memory, storage-class memory, and even AI accelerators like GPUs or field-programmable gate arrays. As a result, the distinction between different layers of the memory hierarchy becomes less rigid, with memory resources becoming more tightly integrated and accessible across the system. This blurring of boundaries offers opportunities for improved performance, energy efficiency, and scalability in modern computing systems, as data-intensive workloads can leverage a more unified and versatile memory architecture. Nevertheless, the fundamental building blocks of such a versatile memory architecture remain upon the underlying memory device technologies. In the following, the mainstream and emerging memory device technologies are surveyed. State of the art from the industry and future trends of these technologies are discussed.

STATIC RANDOM ACCESS MEMORY

Static random-access memory (SRAM) is widely used as the on-chip cache for microprocessors including CPU/GPU and domain-specific accelerators such as tensor processing units. SRAM is still irreplaceable owing to its subnanosecond access speed and unlimited endurance. Depending on the applications, SRAM's bit cell design features high-density or high-performance variants (mainly by sizing the number of fins in the FinFET era). Figure 1 shows the historical scaling trends in the SRAM bit cell area (for the high-density cell) from the planar transistor era to

Digital Object Identifier 10.1109/MC.2024.3363269
Date of current version: 5 April 2024



today's FinFET era. The data points are collected from the industrial reports in leading conferences. The representative microscopic views of the six-transistor bit cell are also shown. As is shown, the SRAM enjoys the scaling benefits of the logic process to the 5-nm/3-nm node, reaching the bit density around 30 Mbit/mm^2. However, the scaling rate has significantly slowed down in recent years. Taking the Taiwan Semiconductor Manufacturing Company's (TSMC's) technology as an example, from 5-nm node to 3-nm node only 5% area reduction is achieved when the high-density bit cell area reduces from 0.021 μm^2 at 5-nm node to 0.0199 μm^2 at 3-nm node.[2] Three-dimensional die stacking of SRAM by advanced packaging techniques, for example, hybrid bonding as used in Advanced Micro Devices' 3D V-Cache,[3] enables the 768-Mbit ultralarge last-level cache for high-performance computing. It is noted that the second generation of 3D V-Cache still used a less advanced 7-nm node for SRAM dies while the processor cores are on a more advanced 5-nm node. The future challenges of SRAM design require innovations in design-technology cooptimization, for example, double/triple layers of wires for wordline/bitline to reduce the parasitic interconnect resistance, backside power rail and power delivery network, stacked nanosheet transistor, folded SRAM bit cell in monolithic 3D integration with complementary field-effect transistor, and so on.

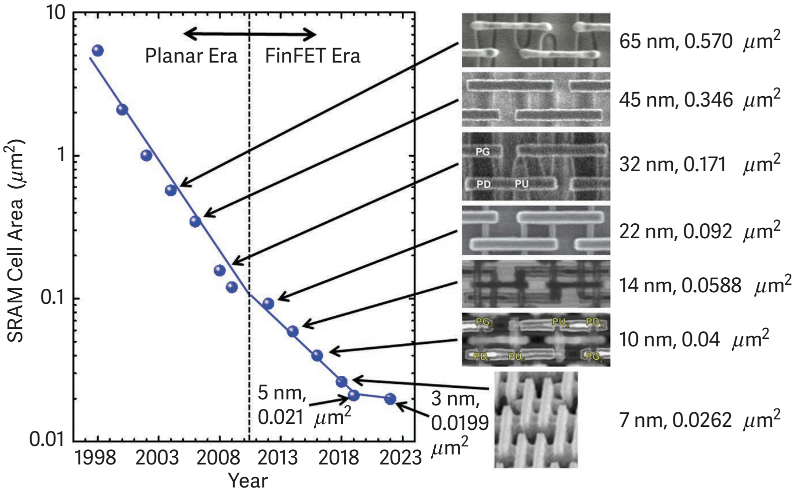


FIGURE 1. Scaling trend of SRAM cell area (for high-density six-transistor bit cell). Adapted from Yu[1] with recent years' data added.

DYNAMIC RANDOM-ACCESS MEMORY

Dynamic random-access memory (DRAM) is used as the main memory, and it is often regarded as off-chip standalone memory with input/output (I/O) links communicating with microprocessors/accelerators. Depending on the applications, DRAM products have different I/O interface protocols such as double data rate (DDR), low power DDR (LPDDR), graphic DDR (GDDR), and high-bandwidth-memory (HBM). For HBM, multiple DRAM dies are stacked vertically with microbump and through-silicon-via and is controlled by the logic base die. High-performance computing platforms are often equipped with GDDR or HBM owing to their ultrafast bandwidth. Figure 2 shows the normalized bit density scaling of various memory technologies, and Figure 3 shows the storage capacity of various memory technologies. DRAM's scaling as of 2023 has reached 12-nm node and the bit density reaches more than 300 Mbit/mm^2 for DDR5[4] and exceed 1 Gbit/mm^2 for HBM3 that employs the 3D die stacking. Extreme ultraviolet lithography and high-k/metal-gate peripheral logic processes have been introduced for DRAM mass production. The future scaling challenges to DRAM include maintaining the sense margin (that is, increasing storage node

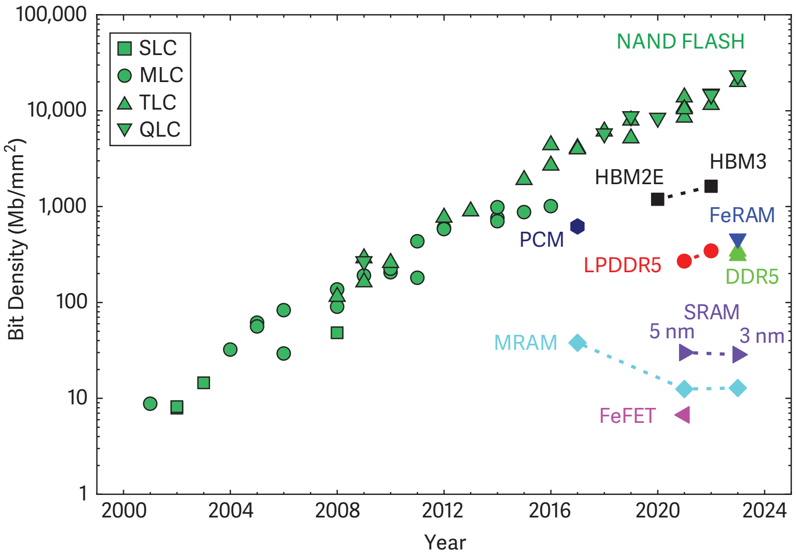


FIGURE 2. Scaling trend of memory bit density for various technologies. Adapted from Yu[1] with recent years' data added. MRAM: magnetic random-access memory.

capacitance by high-aspect ratio stacked capacitors, or reducing the bitline parasitic capacitance), and maintaining the data retention (that is, mitigating the capacitive coupling-induced bit errors such as the row-hammer effect). The possible technological innovations to the next-generation DRAM include employing new channel materials of the cell transistor (for example, amorphous oxide semiconductors) that has intrinsically lower leakage, hiding the peripheral circuits underneath the cell array, or exploiting monolithic 3D stacked DRAM (for example, laying down the DRAM capacitors horizontally or exploiting other mechanisms such as floating-body or avalanche effects for enabling capacitorless bit cells.[5]

NAND FLASH

3D NAND dominates the Flash memory applications in solid-state drives and other mainstream storage media. Today's Flash primarily utilizes the nitride-based charge trap layer in the gate stack as the storage mechanism. 2023 marks the 10th year that the industry transitioned from the 2D NAND architecture to the 3D NAND architecture that takes advantages of the vertical channel in a cost-effective integration solution. State-of-the-art 3D NAND (as of 2023) reaches more than 300 layers and a bit density over 20 Gbit/mm^2.[6] The enabling technologies for realizing such high integration density include the triple-level cell (TLC) or quadruple-level cell (QLC), CMOS under array (CuA), multideck stacking (splitting the vertical channel formation into multiple steps), and so on. The future scaling challenges for 3D NAND include the diminishing sensing current along a very tall vertical channel, the degraded reliability for TLC/QLC operations, and the associated fabrication process difficulties (for example, deep trench etch) toward 1000 layers. The feature directions include possible replacement of the poly-silicon channel materials with higher mobility amorphous oxide semiconductors and possible replacement of the charge-trap layer with a ferroelectric layer in the gate stack for lower program voltage, faster program speed, and improved cycling endurance.[7]

EMERGING MEMORIES

Emerging memories have been extensively explored in the past decade with the hope of supplementing the mainstream technologies (SRAM, DRAM, and NAND Flash) as aforementioned. The tangible applications of emerging memories are mainly serving as embedded nonvolatile memories for the global buffer in microprocessors/accelerators or code storage as in microcontrollers. It is understood that emerging memories are facing difficulties directly competing against the high-density DRAM/NAND products in the standalone memory space. As of 2023, emerging memories are available from foundry platforms at mature legacy nodes. For instance, TSMC is offering resistive random-access memory (RRAM) at 40-nm/28-nm/22-nm nodes.[8] TSMC is also offering spin-transfer torque magnetic random-access memory (STT-MRAM) at 22-nm/16-nm nodes.[9] STMicroelectronics is offering phase change memory (PCM) at 28-nm node,[10] and GlobalFoundries is offering FeFET at 28-nm/22-nm nodes.[11] Sony and Micron are developing FeRAM based on HfO$_2$ material, and the prototype chip density increased from 64 kb[12] to 32 Gb[13] recently. The general characteristics of these emerging memories include sub-100-ns write/read speed, >10^6 endurance cycles, and more years of retention, while the MRAM has a unique advantage of low write voltage (<1 V), and FeFET has a unique advantage of low write energy (<10 fJ/bit). Emerging memories are attractive for certain niche markets, for example, automotive and aerospace electronics where stringent requirements exist on high/low temperature performance or immunity to radiation effects. The challenges for expanding the application space include further lowering the write voltage and making the technologies compatible with more advanced logic processes such as 7 nm or beyond, further improving the endurance and retention and supporting the reliable multilevel operation. The research community is also actively exploring using the emerging memories in the new compute paradigm such as in-memory computing or in-memory search to accelerate the data-intensive workloads such as AI/ML and combinatorial optimization.

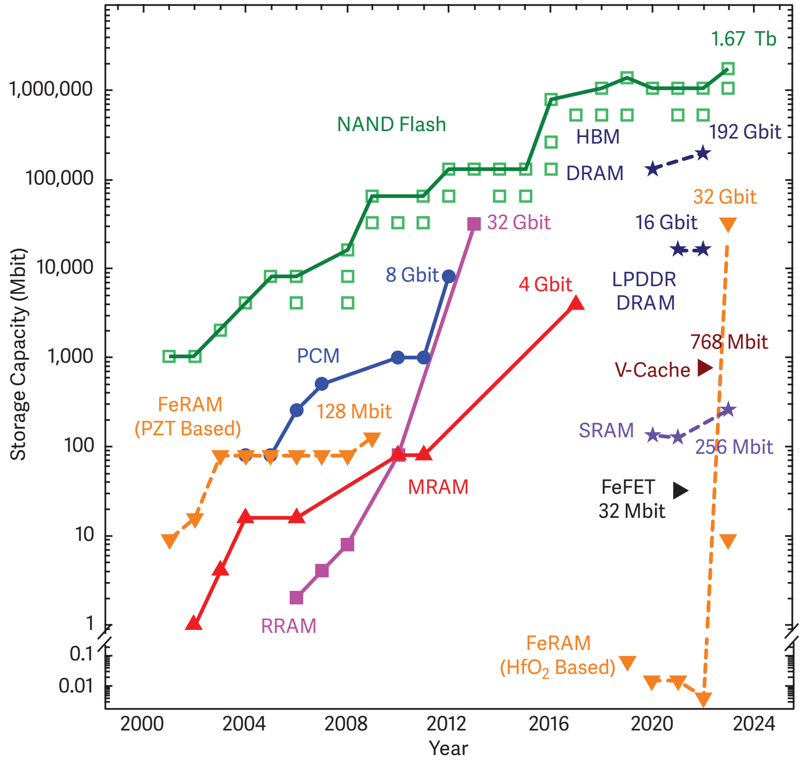


FIGURE 3. Scaling trend of memory capacity for various technologies. Adapted from Yu[1] with recent years' data added. RRAM: resistive random-access memory; FeRAM: ferroelectric random-access memory.

The mainstream memory technologies such as SRAM, DRAM, and NAND Flash have benefited from the technology scaling in the past decades, and the roadmap for continued scaling (with a transition to 3D or even more 3D layers) is defined by the industry. So far, the replacement for these mainstream memory technologies remains elusive. In simple words, SRAM's foreseeable future is better SRAM, DRAM's foreseeable future is better DRAM, and NAND Flash's foreseeable future is better NAND Flash. This is because no other known memory technologies could offer the fast access speed of SRAM while not suffering from endurance degradation. Nor could those technologies provide the high density (thus ultralow cost per bit) of NAND Flash or have a balance between the cost per bit and the access speed/endurance of DRAM. Take the Intel/Micron's 3D XPoint technology[14] (that is based on PCM) as an example of a technology that had its production halted. The business model indicated a high barrier for emerging technology to serve as storage class memory due to competition with high-end NAND Flash based on single-level cell operation which offers relatively fast access speed down to approximately 1 μs. Micron's latest 32-Gbit FeRAM prototype[15] is another example in that it shows superior characteristics that almost meet DRAM specifications while providing certain nonvolatility; however, from the cost perspective, it is still quite challenging for such an emerging technology to gain advantages over mass-produced existing technologies. Therefore, the current role of emerging technologies is to augment mainstream technologies rather than to replace them. New functionalities that are offered by emerging devices and architectures such as in-memory computing or in-memory search will continue to drive further development of these technologies.

REFERENCES

1. S. Yu, *Semiconductor Memory Devices and Circuits*. Boca Raton, FL, USA: CRC Press, 2022.
2. C.-H. Chang et al., "Critical process features enabling aggressive contacted gate pitch scaling for 3nm CMOS technology and beyond," in *Proc. Int. Electron Devices Meeting (IEDM)*, 2022, pp. 27.1.1–27.1.4, doi: 10.1109/IEDM45625.2022.10019565.
3. J. Wuu et al., "3D V-Cache™: The implementation of a hybrid-bonded 64MB stacked cache for a 7nm x86-64 CPU," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2022, pp. 428–429, doi: 10.1109/ISSCC42614.2022.9731565.
4. W. Kim et al., "A 1.1V 16Gb DDR5 DRAM with probabilistic-aggressor tracking, refresh-management functionality, per-row hammer tracking, a multi-step precharge, and corebias modulation for security and reliability enhancement," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2023, pp. 1–3, doi: 10.1109/ISSCC42615.2023.10067805.
5. W.-C. Chen et al., "A 3D stackable DRAM: Capacitor-less three-wordline gate-controlled thyristor (GCT) RAM with >40µA current sensing window, >10^{10} endurance, and 3-second retention at room temperature," in *Proc. Int. Electron Devices Meeting (IEDM)*, 2022, pp. 26.3.1–26.3.4, doi: 10.1109/IEDM45625.2022.10019464.
6. B. Kim et al., "A high-performance 1Tb 3b/cell 3D-NAND flash with a 194MB/s write throughput on over 300 layers i," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2023, pp. 27–29, doi: 10.1109/ISSCC42615.2023.10067666.
7. S. Yoon et al., "QLC programmable 3D ferroelectric NAND Flash memory by memory window expansion using cell stack engineering," in *Proc. IEEE Symp. VLSI Technol. Circuits*, 2023, pp. 1–2, doi: 10.23919/VLSITechnologyand-Cir57934.2023.10185294.
8. C.-X. Xue et al., "A 22nm 4Mb 8bprecision ReRAM computing-in-memory macro with 11.91 to 195.7 TOPS/W for tiny AI edge devices," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2021, pp. 245–247, doi: 10.1109/ISSCC42613.2021.9365769.
9. P.-H. Lee et al., "A 16nm 32Mb embedded STT-MRAM with a 6ns read-access time, a 1M-cycle write endurance, 20-year retention at 150°C and MTJ-OTP solutions for magnetic immunity," in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, 2023, pp. 494–496, doi: 10.1109/ISSCC42615.2023.10067837.
10. F. Arnaud et al., "High density embedded PCM cell in 28nm FDSOI technology for automotive micro-controller applications," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, 2020, pp. 24.2.1–24.2.4, doi: 10.1109/IEDM13553.2020.9371934.
11. S. Muller et al., "Development status of gate-first FeFET technology," in *Proc. IEEE Symp. VLSI Technol. Circuits*, 2021, pp. 1–2.
12. J. Okuno et al., "SoC compatible 1T1C FeRAM memory array based on ferroelectric $Hf_{0.5}Zr_{0.5}O_2$," in *Proc. IEEE Symp. VLSI Technol. Circuits*, 2020, pp. 1–2, doi: 10.1109/VLSITechnology18217.2020.9265063.
13. N. Ramaswamy et al., "NVDRAM: A 32Gbit dual layer 3D stacked non-volatile ferroelectric memory with near-DRAM performance for demanding AI workloads," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, 2023, pp. 1–4.
14. R. Smith. "Intel to wind down Optane memory business—3D XPoint storage tech reaches its end." AnandTech. Accessed: Jul. 28, 2022. [Online]. Available: https://www.anandtech.com/show/17515/intel-to-wind-down-optane-memory-business
15. C. Mellor. "Micron NVDRAM may never become a product." Blocks and Files. Accessed: Jan. 9, 2024. [Online]. Available: https://blocksandfiles.com/2024/01/09/micron-nvdram-might-never-become-a-product/

SHIMENG YU is a professor at the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA. Contact him at shimeng.yu@ece.gatech.edu.

TAE-HYEON KIM is a postdoctoral fellow at the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA. Contact him at thkim@gatech.edu.

EDITOR: Michael Zyda, University of Southern California, zyda@mikezyda.com

This article originally appeared in
Computer
vol. 58, no. 1, 2025

How to Hire a Gen Z Through Gaming

Khizer Khaderi, *Stanford University*
Yusuf Ahmed, *University of Toronto*
Michael Zyda, *University of Southern California*

Given over 3.2 billion people game, including multiple generations of digital natives, we present the opportunity to use advancements in cognitive and perception-based metrology to transform any game into a recruiting tool for Gen Zs.

The use of video games as recruitment tools has evolved significantly since the U.S. Army's launch of *America's Army* in 2002, which pioneered the concept of using a video game to attract potential military recruits. Today, as gaming has become a global phenomenon with more than 3.2 billion players, the opportunity to leverage games beyond entertainment for talent acquisition and skill development has grown. However, fully harnessing gaming as a tool to assess cognitive and perceptual abilities remains untapped (Figure 1).

The Gamer Doctors (TGD), an application developed in part by Dr. Khizer Khaderi and his team at the Stanford Human Perception Lab, represents a transformative step in this direction. TGD allows for the passive capture of both cognitive and perception-based metrologies—the latter being particularly challenging to assess—during gameplay. This innovation provides a more nuanced and scalable way to assess digital natives, such as Gen Z and Gen Alpha, who have developed distinctive cognitive and perceptual skills through their immersive interactions with technology.

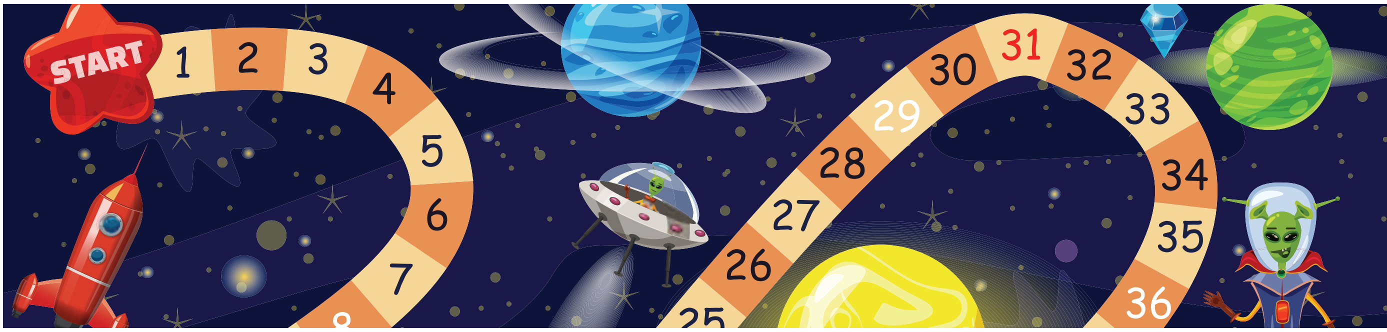
TGD's approach involves creating gamer archetypes, which categorize players based on their cognitive and perceptual profiles, offering insights into their strengths. The original purpose of the gamer archetypes was to personalize game recommendations based on a player's intrinsic skill sets versus traditional methods of measuring a player's gameplay skill sets. During the development of gamer archetypes, we realized the identification of innate skill sets could be applied to job recommendations (Figure 2).

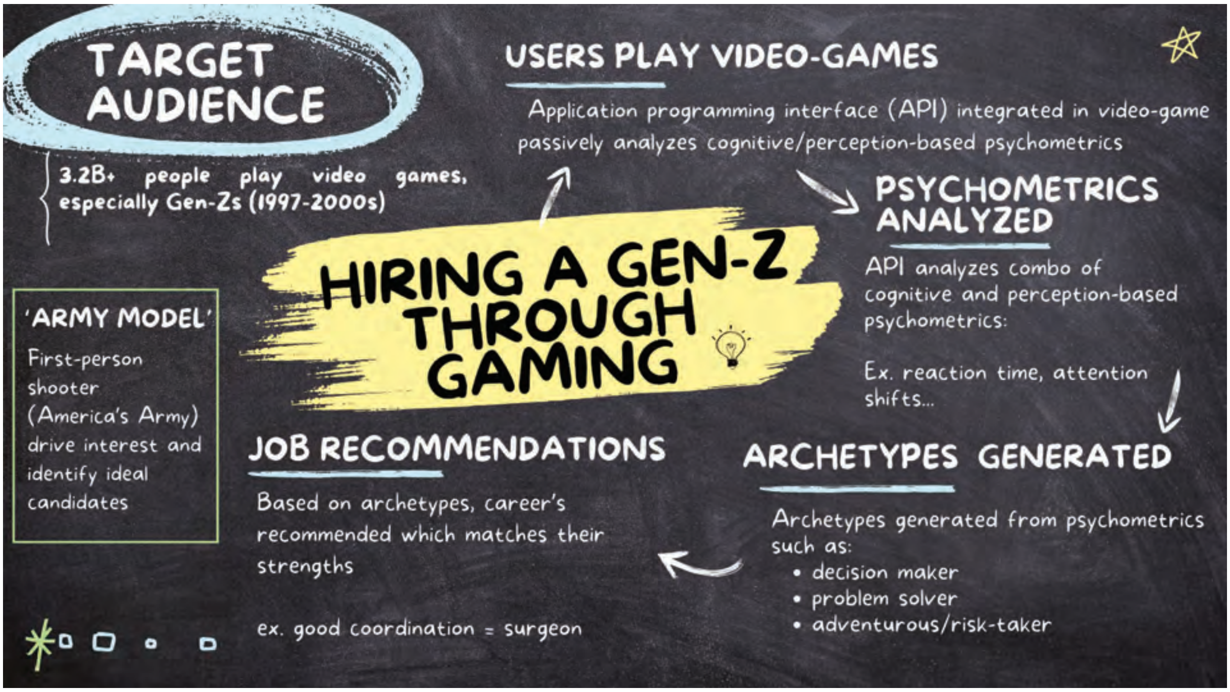
Although inspired by *America's Army*, technology development was not limited to a specific game/genre, but rather focused on development of application programming interfaces (APIs) allowing any video game the capability to capture cognitive and perceptual psychometrics. This strategy democratizes access to psychometric assessments across any game genre but also opens new possibilities for recruitment and workforce development in industries that rely heavily on these skills. By seamlessly integrating assessment tools into any gaming experience, TGD presents an opportunity to transform gaming into a powerful medium for identifying and developing talent in the digital age.

America's Army: How the Army used a first-person shooter to change how we recruit people for jobs

The goal of the *America's Army* project was to build an online 3D PC game that provided the experience of a potential career in the Army. The idea was to make the game as Army-accurate as possible, a game that would educate and engage those young Americans thinking about a potential career in the U.S. Army. The Army was looking for young Americans ages 11 to 14

Digital Object Identifier 10.1109/MC.2024.3485252
Date of current version: 9 January 2025





TARGET AUDIENCE

3.2B+ people play video games, especially Gen-Zs (1997-2000s)

USERS PLAY VIDEO-GAMES

Application programming interface (API) integrated in video-game passively analyzes cognitive/perception-based psychometrics

HIRING A GEN-Z THROUGH GAMING

PSYCHOMETRICS ANALYZED

API analyzes combo of cognitive and perception-based psychometrics:

Ex. reaction time, attention shifts...

'ARMY MODEL'

First-person shooter (America's Army) drive interest and identify ideal candidates

JOB RECOMMENDATIONS

Based on archetypes, career's recommended which matches their strengths

ex. good coordination = surgeon

ARCHETYPES GENERATED

Archetypes generated from psychometrics such as:

- decision maker
- problem solver
- adventurous/risk-taker

FIGURE 1. How to hire a Gen Z through gaming.

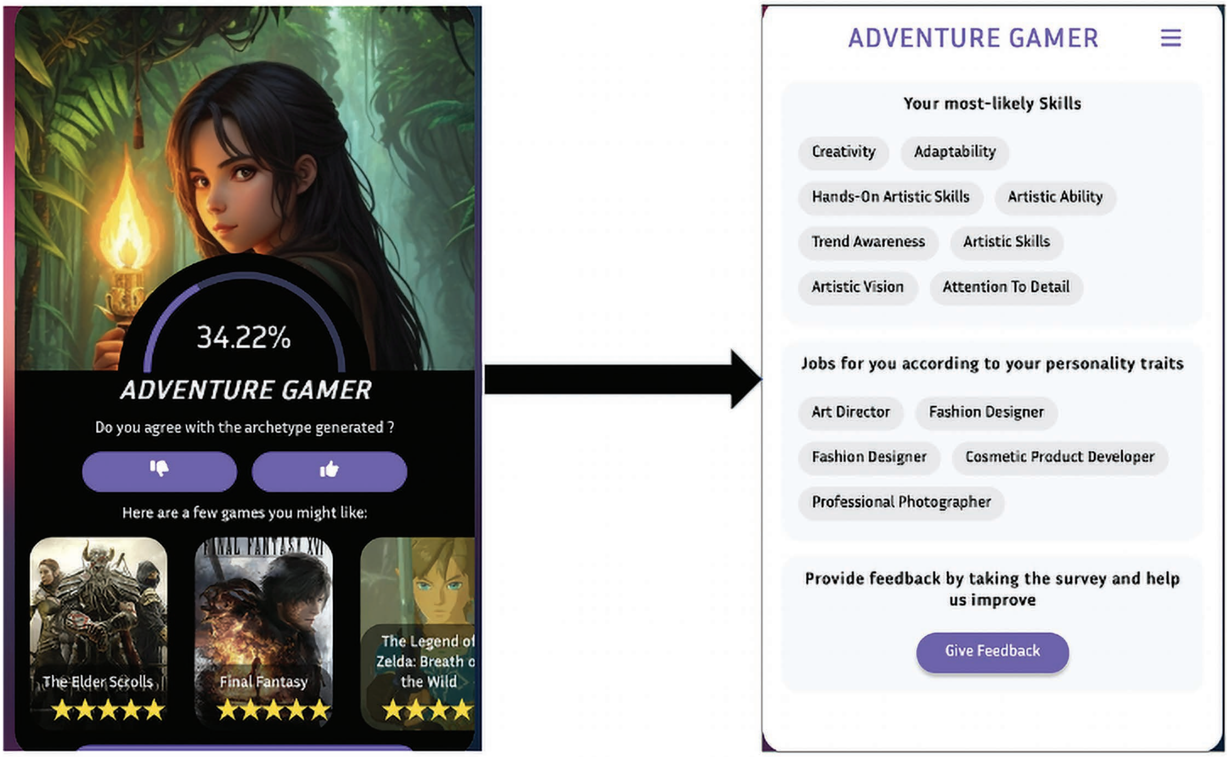
to play this game. The Army knew if a young American played this game between those ages, then when they turned 18 they would be twice as likely to consider a career in the Army as young Americans who knew nothing about the Army. The project was pretty important as the Army had failed to meet its recruiting goals in 1999—the *America's Army* Project became the point of the spear for recruiting for the U.S. Army once the game was released.[1,2]

The idea for the game came out of an observation the Army made of its recruits. New recruits to the Army typically had an Army toy, a GI Joe, or a model tank, in their kit in the barracks. When the Army asked about the origin of the toys, they found that someone in their family had given it to them sometime during their ages of 11–14. The Army discovered that young Americans with such toys were twice as likely to consider a career in the Army than those without.

The Army wondered if an online game could serve the same recruiting function and decided to fund the development of that game.

> *THE IDEA WAS TO MAKE THE GAME AS ARMY-ACCURATE AS POSSIBLE, A GAME THAT WOULD EDUCATE AND ENGAGE THOSE YOUNG AMERICANS THINKING ABOUT A POTENTIAL CAREER IN THE U.S. ARMY.*

As part of that game's development, the Army additionally studied the issue of whether the Armed Services Vocational Aptitude Battery (ASVAB) score could be computed from game play. The ASVAB score project revealed that it could be computed from



34.22%

ADVENTURE GAMER

Do you agree with the archetype generated ?

Here are a few games you might like:

The Elder Scrolls ★★★★★

Final Fantasy ★★★★★

The Legend of Zelda: Breath o the Wild ★★★★

ADVENTURE GAMER

Your most-likely Skills

Creativity · Adaptability · Hands-On Artistic Skills · Artistic Ability · Trend Awareness · Artistic Skills · Artistic Vision · Attention To Detail

Jobs for you according to your personality traits

Art Director · Fashion Designer · Fashion Designer · Cosmetic Product Developer · Professional Photographer

Provide feedback by taking the survey and help us improve

Give Feedback

FIGURE 2. Gamer archetype identifying skills and recommendations for games and jobs.

properly designed game play. The Army decided not to ship the ASVAB score part of the *America's Army* game due to privacy concerns.

The ASVAB score is used by the Army to determine potential career paths for its recruits. The score is used to bucketize recruits into potential careers in which they will do well. The Army, like all employers, wants everyone it hires to succeed, and one of the best ways is to channel new recruits into appropriate career paths.

The *America's Army* game turned out to be the most successful recruiting tool ever developed by the Army and it ran online from July 2002 to February 2022. Following the closure of the online game, the Army began, once again, missing its recruiting goals, and general officers began once again asking if the game could be restarted.

EVOLVING GAME-BASED RECRUITING FOR DIGITAL NATIVES

The recruitment of digital natives—those born in an age defined by ubiquitous digital technology—has proven to be a unique challenge for traditional recruitment models. *America's Army* was among the first initiatives to blend gaming with recruitment, using a military simulation game to attract and engage potential recruits.[1] However, broader adoption of games as recruitment tools across industries has lagged behind.

Digital natives, including Gen Z and Gen Alpha, exhibit distinct cognitive characteristics, such as enhanced multitasking abilities and increased visual processing speeds, due to their exposure to fast-paced digital environments like video games.[4] Leveraging gaming platforms to passively capture these abilities is where technology such as TGD becomes critical. TGD assesses both cognitive and perceptual skills—providing insights into candidates' problem-solving abilities, attention, and reaction times without the need for traditional testing environments.[5]

PERCEPTION AND COGNITIVE MODELS FROM GAMEPLAY

Recent advances in cognitive and perceptual sciences have emphasized the role of perception in game-based environments. Studies show that perceptual abilities, such as depth perception, reaction time, and spatial

FIGURE 3. Visual function, digital behavior, and the vision performance index.

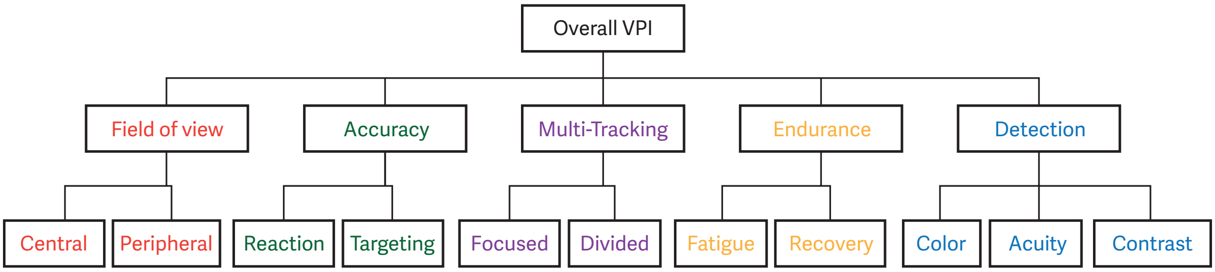


FIGURE 4. *America's Army,* the Pentagon's video game.[3]

awareness, can be measured and improved through gameplay.[4,5,6,7,8] These abilities are critical in professions requiring quick, accurate decisions, such as health care, aviation, and cybersecurity.

TGD uses cognitive and perceptual models, derived from psychometric metrology, to capture how players behave in digital environments. These models identify key abilities, such as decision-making and spatial reasoning, which are indicative of job success.[10,11] By analyzing in-game behaviors, such as reaction times or attention shifts, the system provides a passive methodology to comprehensive understanding of the player's cognitive and perceptual profile.

This approach aligns with research showing that gaming environments provide valid simulations of real-world cognitive tasks, including problem-solving and multitasking.[12] Studies also show that perceptual skills developed during gameplay are transferable to tasks that require high levels of spatial awareness and visual acuity, such as surgery or piloting.[13]

Prior to developing the TGD application, Dr. Khaderi's work in perception-based psychometrics included developing the vision performance index (VPI) to measure key cognitive and perceptual functions through interactive media.[9] This tool uses games to capture fields such as visual attention, field of view, multitasking, and endurance (Figure 3).

In constructing the VPI, attention was turned to creating simple game experiences, with retro-style designed games, as noted in Figure 4. As a player goes through the game experience, real time passive capture of raw psychometric signals are occurring. The player receives both an in-game score, as well as a VPI score, with the breakdown of the VPI score in five dimensions noted in Figure 5.

> *THE ARMY, LIKE ALL EMPLOYERS, WANTS EVERYONE IT HIRES TO SUCCEED, AND ONE OF THE BEST WAYS IS TO CHANNEL NEW RECRUITS INTO APPROPRIATE CAREER PATHS.*

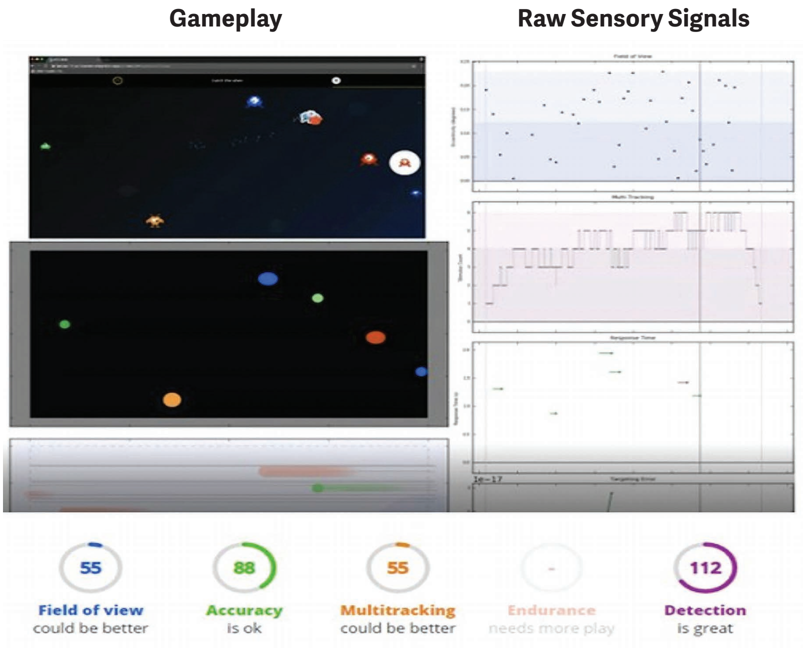


FIGURE 5. Passive capture of psychometrics through games and the vision performance index.

The Stanford Human Perception Lab (HPL) conducted studies of esport professionals across different genres of games utilizing the VPI. Given the demanding nature of training of esport professionals, identifying methods of optimizing performance, while balancing a healthy lifestyle was an important objective of this study. The study revealed the VPI of players who worked out at least 2 h a week and trained in intervals of 2–5 h of screen time were able to maintain high levels of success in competition (Figure 6).

STUDIES SHOW THAT PERCEPTUAL ABILITIES, SUCH AS DEPTH PERCEPTION, REACTION TIME, AND SPATIAL AWARENESS, CAN BE MEASURED AND IMPROVED THROUGH GAMEPLAY.

By tracking in-game actions and reactions, VPI offers a noninvasive way to assess abilities critical for decision-making and task execution in high-performance environments.[7] Khaderi's work on VPI has been utilized in the field of sports and esports to measure player performance, most recently endorsed by the International Olympic Committee.[16]

Khaderi's research demonstrates that perception-based assessments are effective at identifying changes in cognitive and visual perception performance, as well as laying the groundwork for applying this metrology in other contexts, such as recruitment to predict an individual's performance in complex, high-pressure jobs.[7] TGD's platform integrates these cognitive and perceptual sciences by capturing how players interact with digital environments, transforming these interactions into measurable outputs.

Gaming as the future of Human Resources

The science behind TGD's platform is grounded in cognitive psychology and psychometrics, which have long been used to predict job performance based on task engagement.[7,8,9,10,11,12] Video games provide rich datasets, allowing researchers to capture how players make decisions, react to stimuli, and manage tasks, all of which are important indicators of professional potential.[11,12,13,14,15]

The system's models are informed by decades of research in perception science and psychophysics, such as Khaderi's VPI studies, which demonstrated that gameplay could reveal key cognitive and visual perception functions in healthy and vision-impaired populations.[7]

Building on the work of Dr. Khaderi and team, the TGD application was designed with Gen Z in mind. Given the majority of digital natives play games for entertainment, the skills they have passively developed in simulated environments have demonstrated the potential to be translatable in the real world.[4,5,6]

Game-based environments allow for the recognition of skills often overlooked in traditional HR assessments. For example, skills like hand-eye coordination, reaction speed, and spatial memory can be objectively measured through gameplay,

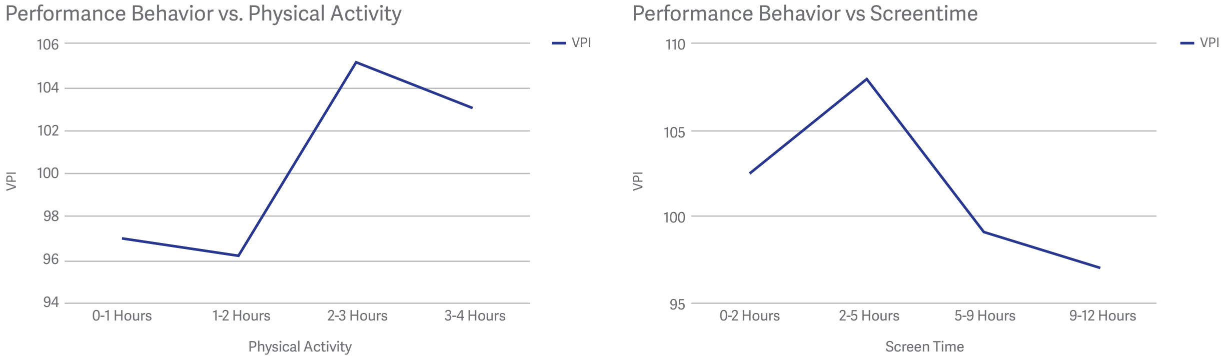


FIGURE 6. Lifestyle factor effects on esport professional performance.

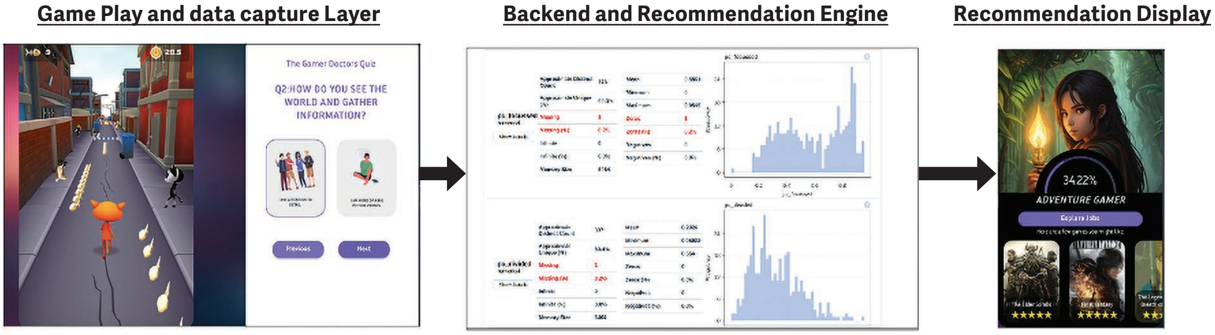
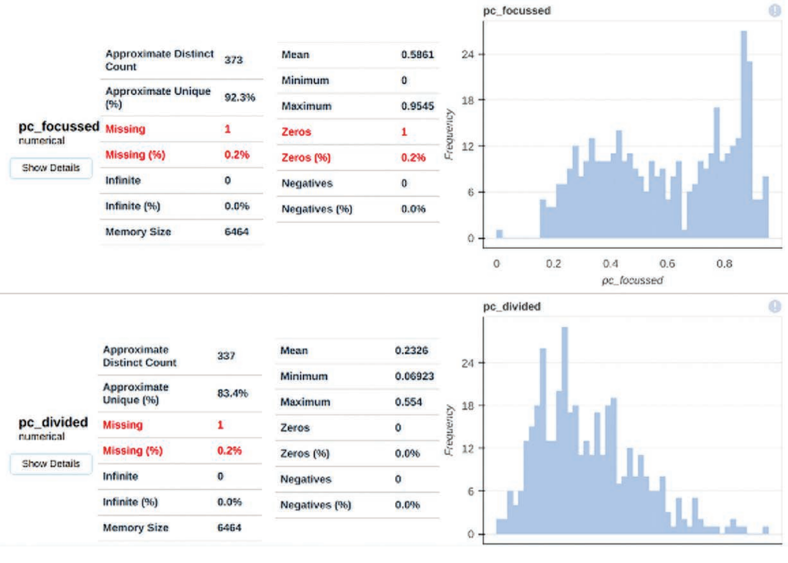

Game Play and data capture Layer
Backend and Recommendation Engine
Recommendation Display
The Gamer Doctors Quiz
Q2: HOW DO YOU SEE THE WORLD AND GATHER INFORMATION?
Previous
Next
34.22%
ADVENTURE GAMER

FIGURE 7. TGD workflow.

providing a reliable assessment of a candidate's capabilities.[14] TGD captures these skills and offers a detailed profile of the player's strengths and weaknesses, supported by research showing that gaming assessments can predict job performance in fields such as logistics and engineering.[15]

Similar to the methodology used in measuring player performance with the VPI, TGD captures cognitive and perception-based psychometrics passively as shown in the infinite runner game displayed in Figure 7. Real-time passive measurement of player multitasking skill is noted in the workflow in Figure 7, as well as in more detail in Figure 8. The resulting recommendation involves modeling of the psychometrics into a gamer archetype as noted in Figure 9.



pc_focussed
numerical

Show Details

Approximate Distinct Count	373	Mean	0.5861
Approximate Unique (%)	92.3%	Minimum	0
Missing	1	Maximum	0.9545
Missing (%)	0.2%	Zeros	1
Infinite	0	Zeros (%)	0.2%
Infinite (%)	0.0%	Negatives	0
Memory Size	6464	Negatives (%)	0.0%

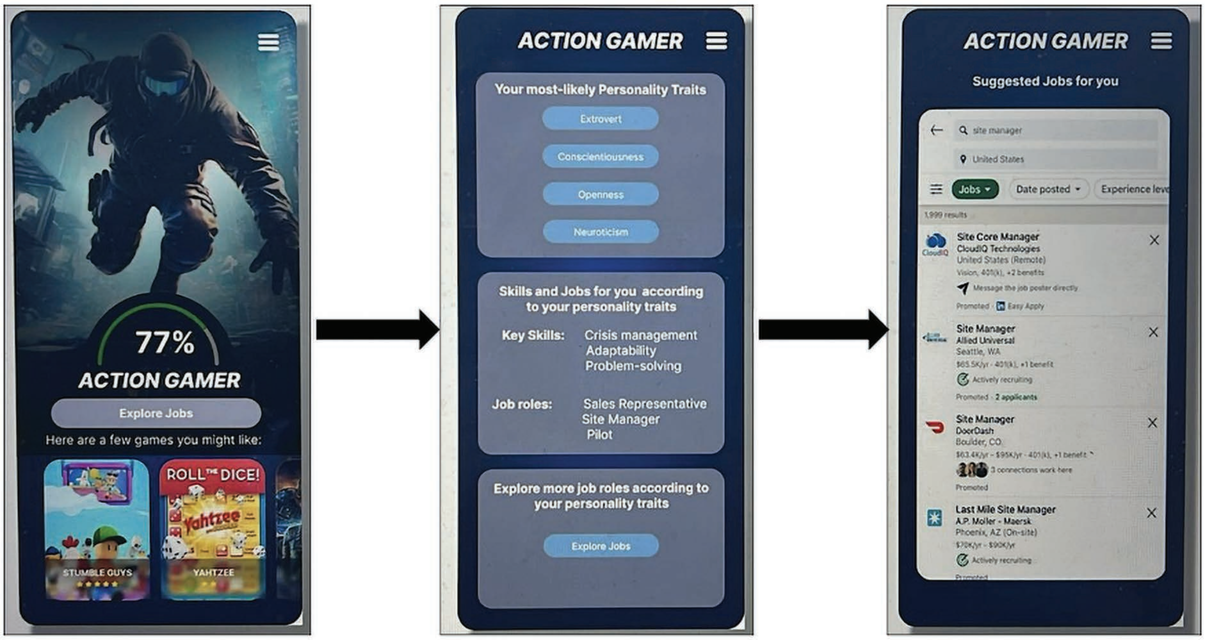
pc_divided
numerical

Show Details

Approximate Distinct Count	337	Mean	0.2326
Approximate Unique (%)	83.4%	Minimum	0.06923
Missing	1	Maximum	0.554
Missing (%)	0.2%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Negatives	0
Memory Size	6464	Negatives (%)	0.0%

FIGURE 8. TGD real-time passive capture of player multi-tasking capabilities.

TGD's API allows for seamless integration with any game, regardless of genre or platform. This flexibility ensures that game developers and recruiters alike can



77%

ACTION GAMER

Here are a few games you might like:

ACTION GAMER

Your most-likely Personality Traits

Skills and Jobs for you according to your personality traits

Key Skills: Crisis management, Adaptability, Problem-solving

Job roles: Sales Representative, Site Manager, Pilot

Explore more job roles according to your personality traits

ACTION GAMER

Suggested Jobs for you

site manager

United States

1,999 results

Site Core Manager
CloudIQ Technologies
United States (Remote)
Vision, 401(k), +2 benefits
Message the job poster directly
Promoted · Easy Apply

Site Manager
Allied Universal
Seattle, WA
$65.5K/yr · 401(k), +1 benefit
Actively recruiting
Promoted · 2 applicants

Site Manager
DoorDash
Boulder, CO
$63.4K/yr - $95K/yr · 401(k), +1 benefit
3 connections work here
Promoted

Last Mile Site Manager
A.P. Moller - Maersk
Phoenix, AZ (On-site)
$70K/yr - $90K/yr
Actively recruiting
Promoted

FIGURE 9. Game skills to job site mapping.

leverage TGD's psychometric assessments to evaluate players' abilities. The API can capture data from a wide range of games, offering valuable insights into cognitive and perceptual skills across diverse contexts and simulated environments.[7,8,9]

Once player archetypes are generated, TGD suggests career paths based on their innate and game play skills. Research shows that individuals' performance in tasks like gaming correlates with their ability to succeed in professional domains requiring similar competencies.[11,12,13,14,15] For example, players excelling in puzzle-solving games may have strong analytical capabilities suited for careers in data science or software development, while those who excel in coordination-heavy games may be more suited for hands-on, precision-based professions.[11,12,13,14,15]

Game-based psychometric assessments represent a significant advancement and opportunity for human resources. By providing a more nuanced understanding of candidates' abilities, TGD's platform allows human resource (HR) professionals to move beyond traditional assessments and gain insights into a candidate's cognitive and perceptual capabilities, aligning recruitment strategies with the digital habits of Gen Z and Gen Alpha.[10,13] These developments reflect broader trends in HR, where AI and machine learning are increasingly used to optimize talent acquisition and development.

INITIAL RESULTS AND POTENTIAL FUTURES

TGD's initial results have shown that its system effectively identifies players' strengths and suggests appropriate career paths. However, challenges remain, such as addressing cultural differences in gameplay styles, which can impact how psychometric data are interpreted across populations. Furthermore, the reliance on gaming as an assessment tool may introduce bias against those unfamiliar with specific game genres, limiting the system's applicability to nongaming populations.

Looking forward, future iterations of TGD could integrate more comprehensive assessments, including emotional and social intelligence, to provide a holistic understanding of a candidate's capabilities. As game-based recruitment tools evolve, they could become key components of a data-driven, personalized recruitment ecosystem.

The evolution of game-based recruitment has made a significant leap from *America's Army*, which pioneered the use of a single, purpose-built video game for military recruitment. While *America's Army* successfully demonstrated the potential of gaming to attract and engage talent, platforms like TGD represent a major advancement by enabling the passive capture of cognitive and perceptual data from *any* video game. This flexibility is revolutionary, expanding the scope of game-based recruitment beyond a single environment to reach millions of gamers across genres and platforms.

TGD's ability to integrate seamlessly with any game allows recruitment to assess a far wider range of skills, capturing diverse cognitive and perceptual abilities across various gameplay scenarios. Whether players are solving complex puzzles, making rapid strategic decisions, or demonstrating precision in action-based tasks, TGD identifies key traits and matches them to relevant career paths. This shift democratizes the recruitment process, making it applicable not just to the military but to industries across the board that seek to leverage the talent of digital natives like Gen Z and Gen Alpha.

By unlocking the potential of any video game as a recruitment platform, TGD heralds the future of recruitment—one that is inclusive, scalable, and tailored to the unique skills of digital natives. As gaming becomes a dominant part of everyday life for younger generations, these tools will transform how we identify and nurture talent, mirroring the foundational success of *America's Army* but vastly expanding its impact and applicability across diverse sectors. Through game-based assessments, we stand on the cusp of revolutionizing talent acquisition, fostering a more personalized and data-driven approach to developing the workforce of tomorrow.

COMMENTS?

If you have comments about this article, or topics or references I should have cited or you want to rant back to me on why what I say is nonsense, I want to hear. Every time we finish one of these columns, and it goes to print, what I'm going to do is get it up online and maybe point to it at my Facebook (mikezyda) and my LinkedIn (mikezyda) pages so that I can receive comments from you. Maybe we'll react to some of those comments in future columns or online to enlighten you in real time! This is the "Games" column. You have a wonderful day.

ACKNOWLEDGMENT

We wish to acknowledge all the contributing members of the Stanford Human Perception Laboratory and Krutin Shah in particular for getting this paper out on time.

REFERENCES

1. R. Nichols, *America's Army and the Recruitment and Marketing of Games*. Durham, NC, USA: U.S. Army Research Office, 2010.
2. M. Zyda, "Weapons of mass distraction - the *America's Army* Game at 20," *Computer*, vol. 55, no. 7, pp. 112–122, Jul. 2022, doi: 10.1109/MC.2022.3169388.
3. Steam Powered. "America's army: Proving grounds." Steam. Accessed: Nov. 8, 2024. [Online]. Available: https://store.steampowered.com/app/203290/Americas_Army_Proving_Grounds/
4. L. M. Carrier, N. A. Cheever, L. D. Rosen, S. Benitez, and J. Chang, "Multitasking across generations: Multitasking choices and difficulty ratings in three generations of Americans," *Comput. Hum. Behav.*, vol. 25, no. 2, pp. 483–489, 2009, doi: 10.1016/j.chb.2008.10.012.
5. V. J. Shute and M. Ventura, "The power of play: Designing video games that foster learning," in *Trends and Issues in Instructional Design and Technology*, New York, NY, USA: Routledge, 2013.
6. C. S. Green and D. Bavelier, "Action video game modifies visual selective attention," *Nature*, vol. 423, no. 6939, pp. 534–537, 2003, doi: 10.1038/nature01647.
7. D. Bavelier, C. S. Green, D. H. Han, P. F. Renshaw, M. M. Merzenich, and D. A. Gentile, "Brains on video games," *Nature Rev. Neurosci.*, vol. 12, no. 12, pp. 763–768, 2011, doi: 10.1038/nrn3135.
8. I. Spence and J. Feng, "Video games and spatial cognition," *Rev. Gen. Psychol.*, vol. 14, no. 2, pp. 92–104, 2010, doi: 10.1037/a0019491.
9. Y. Ahmed et al., "Democratizing healthcare in the Metaverse: How video games can monitor eye conditions using the vision performance index: A pilot study," *Ophthalmol. Sci.*, vol. 4, no. 1, 2023, Art no. 100349, doi: 10.1016/j.xops.2023.100349.

10. R. E. Mayer, *The Cambridge Handbook of Multimedia Learning*. Cambridge, U.K.: Cambridge Univ. Press, 2019.
11. J. P. Gee, *What Video Games Have to Teach Us About Learning and Literacy*. New York, NY. USA: Palgrave Macmillan, 2007.
12. E. Adams, *Fundamentals of Game Design*. Berkeley, CA, USA: New Riders, 2010.
13. I. Granic, A. Lobel, and R. C. M. E. Engels, "The benefits of playing video games," *Amer. Psychologist*, vol. 69, no. 1, pp. 66–78, 2014, doi: 10.1037/a0034857.
14. J. P. Gee, "Learning by design: Games as learning machines," *Interactive Educ. Multimedia*, vol. 2, no. 1, pp. 15–23, 2005, doi: 10.2304/elea.2005.2.1.5.
15. J. Gackenbach, *Video Game Play and Consciousness*. Commack, NY, USA: Nova (Science Publishers), 2012.
16. M. C. Moe et al., "International Olympic Committee (IOC) consensus paper on sports-related ophthalmology issues in elite sports," *BMJ Open Sport Exercise Med.*, vol. 9, no. 3, 2023, Art no. e001644, doi: 10.1136/bmjsem-2023-001644.

KHIZER KHADERI is a clinical associate professor at the Byers Eye Institute and the founding director of the Stanford Human Perception Laboratory and Vision Performance Center, Stanford University, Palo Alto, CA 94303, USA. Contact him at kkhaderi@stanford.edu.

YUSUF AHMED is a PGY3 resident physician in ophthalmology at the Department of Ophthalmology & Vision Sciences University of Toronto, Toronto, ON M5T 3A9, Canada. Contact him at y.ahmed@mail.utoronto.ca.

MICHAEL ZYDA is the founding director of the Computer Science Games Program and a professor emeritus of engineering practice in the Department of Computer Science, University of Southern California, Los Angeles, CA 90089 USA. Contact him at zyda@mikezyda.com.

CALL FOR SPECIAL ISSUE PROPOSALS

Computer solicits special issue proposals from leaders and experts within a broad range of computing communities. Proposed themes/issues should address important and timely topics that will be of broad interest to *Computer*'s readership. Special issues are an essential feature of *Computer*, as they deliver compelling research insights and perspectives on new and established technologies and computing strategies.

Please send us your high-quality proposals for the 2025–2026 editorial calendar. Of particular interest are proposals centered on:

- 3D printing
- Robotics
- LLMs
- AI safety
- Dis/Misinformation
- Legacy software
- Microelectronics

Proposal guidelines are available at:
www.computer.org/csdl/magazine/co/write-for-us/15911

IEEE COMPUTER SOCIETY

IEEE

EDITOR: Markus Borg, CodeScene, markus.borg@codescene.com

DEPARTMENT: REQUIREMENTS

This article originally appeared in IEEE Software vol. 40, no. 5, 2023

My REvelation: Unveiling an Unseen Career in Requirements

Sofija Hotomski

FROM THE EDITOR

The theme of this issue is "developing your software engineering career"—a boundless topic as software continues to shape our world and society with every passing moment. What could a career entail in the field of requirements engineering? We are privileged to present the insights of an esteemed professional who serendipitously embarked on an RE journey. Some of our readers might have met her at conferences. Now you get the chance to learn what she's been up to since defending her Ph.D. in 2019!—*Markus Borg*

One day in October 2014, I received an e-mail from Boris Spasojevic, one of my best friends. He was doing his Ph.D. in Switzerland at the University of Bern. The e-mail subject was, "You might be interested or maybe not," and the content was just a link to an open Ph.D. position in RERG, the requirements engineering (RE) research group at the University of Zurich, led by Prof. Martin Glinz. Boris had met Irina Koitz, a RERG member at the time, at a conference. He learned about the position from her and realized it was well-suited for me. I didn't. Therefore, my immediate response to him was "I am probably not interested." After all, I already had a good job that I really enjoyed.

SHORT E-MAIL, MAJOR CAREER MOVE

However, after rereading the project description, I got incredibly intrigued. The research project was about automated updates of software documentation. I did indeed struggle a lot with outdated documents at work. Thus, I applied for the position and was offered the job. Signing the contract was how it all "formally" started—although, as previously stated, my RE career was already well on its way thanks to my job at Schneider Electric.

After my move to Zürich, I learned that my position involved both work on my thesis and teaching duties. I would be a teaching assistant for software engineering (no problem!) and RE (wait, what?). As the start of the RE course drew closer, my panic increased. I didn't know what RE was. How could I possibly teach it to the students? Luckily, reading the course materials reassured me. I saw that 80% of the content I had either learned during my university studies or had a solid understanding of from my previous job at Schneider Electric. RE had been with me all the time!

DISCOVERING RE

Careful use of terminology matters. Many software professionals are unknowingly working with RE tasks. As a result, the RE career path might be unknown to them. They would probably not apply for a job opening explicitly mentioning RE. Speaking for myself about

Digital Object Identifier 10.1109/MS.2023.3283830
Date of current version: 13 October 2023

terminology discrepancies and open doors, I was amazed at how many openings there are for "Requirements Engineers" in the job market! Many of us learned RE topics as part of our engineering studies without ever knowing the term used in industry. How could I have missed this?

First, I thought RE was a new term. However, I quickly found out that Martin Glinz had established his Requirements Engineering Research Group in Zurich back in 1993 and that he had been teaching a dedicated RE course for many years. I also discovered that RE as a discipline was about 40 years old. I was simply oblivious to it.

After finishing my Ph.D. in 2019, I started looking for jobs. Luckily, this time I knew what terms to search for! I started working as a requirements engineer for ASMIQ, a daughter company of the Swiss Post, the national postal service of Switzerland. In this position, my responsibilities went beyond pure RE, and I gained more experience in product management and decided to pursue a career in that direction. Now I work as a product manager for cloud applications in the fire safety domain at Siemens. Still engaging in a lot of RE tasks—and I love every second of it!

Working as a requirements engineer and product manager over the last couple of years, I gained new experiences in requirements management and communication. In the remainder of this column, I share two challenges for which there are still no magic solutions. To end on a positive tone, I also present a requirements-related practice that I've found helpful for defining requirements and communicating them to the development team.

DEPENDENT FRAGMENTS, MONSTERS, AND SILOED REPOS

The first challenge is what I call the "I-S" contradiction. I've encountered this in both the small and big companies I've worked for. Are you familiar with the INVEST concept from the agile community?[1] It states that a good user story (a requirement) should be both independent (I) and small (S). This is particularly tricky in complex projects that combine software and hardware. If a story is small, it usually does not contain all the necessary pieces of information needed to specify a reasonable part of the functionality. Such small user stories often depend on others that complete the functionality or enable it—they are far from independent.

IF A STORY IS SMALL, IT USUALLY DOES NOT CONTAIN ALL THE NECESSARY PIECES OF INFORMATION NEEDED TO SPECIFY A REASONABLE PART OF THE FUNCTIONALITY.

On the other hand, there are also what I call "monster requirements." Analogous to the monster class antipattern in object-oriented programming,[2] also known as God classes, they contain a lot of responsibilities. A monster requirement captures all necessary and unnecessary details and, therefore, is difficult to comprehend. For instance, some engineers tend to write down entire discussions—over several meetings about the topic—as part of the requirement description. This can make the requirement unnecessarily massive to digest. Especially if you were not part of the earlier discussions and simply seek a concise written description—without needing to call for an additional meeting, as your agenda is already full.

Nevertheless, such monster requirements often cover everything needed without depending on other requirements. I often wonder what the best tradeoff is when the INVEST principle is challenged. Or is it better to accept either INVET or NVEST in such cases?

The second challenge comes from my experience working in ASMIQ, a medium-sized startup company.

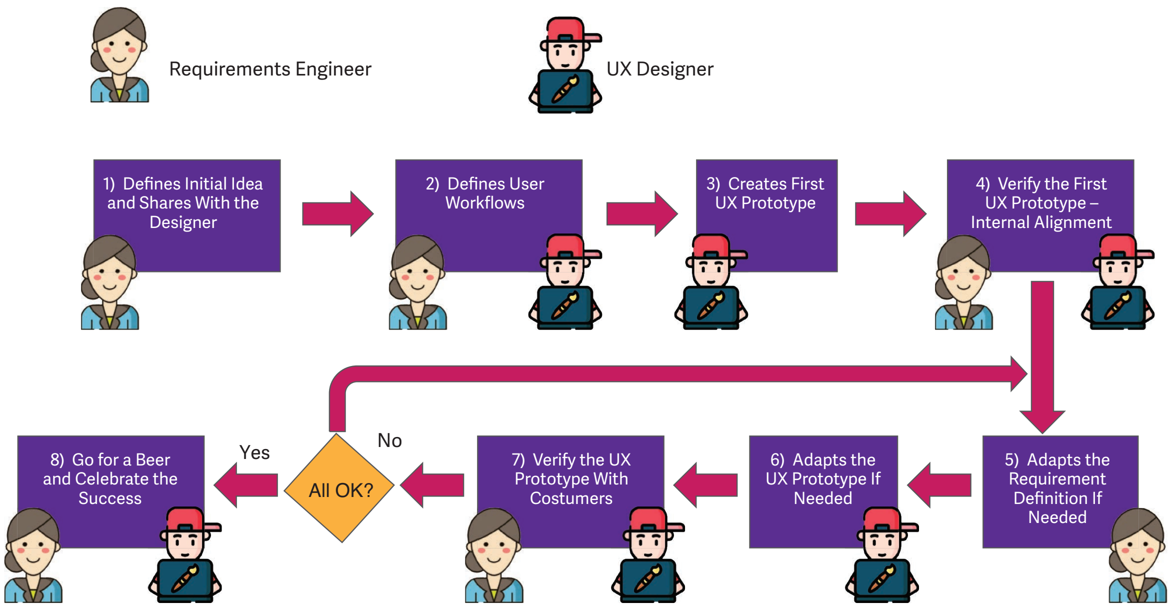

Requirements Engineer
UX Designer
1) Defines Initial Idea and Shares With the Designer
2) Defines User Workflows
3) Creates First UX Prototype
4) Verify the First UX Prototype – Internal Alignment
5) Adapts the Requirement Definition If Needed
6) Adapts the UX Prototype If Needed
7) Verify the UX Prototype With Costumers
All OK?
No
Yes
8) Go for a Beer and Celebrate the Success

FIGURE 1. Joint requirements definition with a UX designer.

The company relied on a polyrepo organization, that is, they used multiple repositories for the source code management version control system.[3] The development team wanted me to write user stories to ensure that each could be implemented in a single repository. From their perspective, writing a user story describing functionality that required source code changes in several repositories was bad practice. I spent hours and hours explaining that a user story describes a piece of business functionality. As a requirements engineer, I do not need to know the internals of the system architecture. My focus was on writing good, understandable requirements with acceptance criteria that indicate whether the story had been implemented (correctly) or not. However, the local implementation of the DevOps process required that a piece of code in a repository must solve a particular story. Cross-repository user stories violated the process. This was quite the straitjacket for me as the requirements engineer.

A PROVEN-IN-USE RE PROCESS

After a decade in professional RE roles, I can also share a helpful approach related to communication. When defining requirements, I have often collaborated extensively with user experience (UX) designers. To prepare a complete and easily understandable requirement, this collaboration really needs to be close and continuous. Especially with tight deadlines, working in silos will never lead to good results. Regardless of the development context—big or small project, big or small company, agile development or not—I've found a collaborative process with UX designers that work really well for defining user-facing requirements. Figure 1 shows an overview as follows:

1. *Defining the initial idea and sharing it with a UX designer*: First, I define the initial idea for a feature. This idea can come as a direct request from customers or originate from internal RE activities. In the latter case, it should be based on knowledge about the system and potential user benefits. Once the idea is mature and documented, I share it with the UX designer and explain the main idea and how it should work.
2. *Defining user workflows in detail*: Together, the UX designer and I define how the users shall use the functionality. We define concrete steps and application workflows. The workflows contain the list of all actions that the users can do and the content they will see when performing those actions. In short, "when I click here, this happens."

3. *Creating the first prototype*: The UX designer creates a first prototype, which is far from perfect but serves as the visualization of the initial idea.
4. *Verifying the first prototype internally*: Before going to the customers or other stakeholders, the UX designer and I discuss the prototype and verify that we have the same understanding of the functionality. We also ensure that the design provides an adequate form for the given functionality.

5–6. *Adapting the prototype and/or the requirement if needed*: In the previous step, we sometimes discover that the requirement should be adjusted or that the design should be changed. This happens, for instance, when the UX designer or I have a better idea of how to represent data or when the idea evolves during the process.

7. *Verify the prototype with customers*: Ideally, I have good contacts with future users. If this is the case, I approach some of them for an early validation of the features using the prototype. Prototypes are much easier to validate than requirements artifacts because they nicely demonstrate the functionality. The users can then imagine what the result will look like.[4] If needed, we return to Steps 5—6 to address the customers' feedback.
8. *Celebrate the success*: If the customer is happy, the UX designer and I make sure to celebrate the success together! No matter if we're co-located or working physically together, we take the time to recognize our good results. This is great for team building and a good basis for future work together!

I've found this process useful again and again. In cases where I previously struggled in my RE role, I keep returning to this process as it helps me meet three core expectations of my work. First, the process helps me keep deadlines for requirements definitions thanks to early feedback. Second, the resulting requirements contain enough detail for the developers to implement the features. Third, it greatly increases the chance that the users get the features they want.

To conclude, there will always be challenges in defining and communicating clear, understandable requirements. Also, requirements engineers must learn to navigate the existing processes in their respective organizations. On the one hand, we know that we need to adapt our RE practices according to the specifics of each company. On the other hand, the RE community has collected a toolbox of techniques that can be adapted for various contexts. Looking back, I find that my early career in RE has been highly fulfilling. This was just the beginning, but I hope my story can inspire others to pursue similar career paths!

REFERENCES

1. B. Wake, "INVEST in good stories, and SMART tasks," XP123, Bellevue, NE, USA, 2003. [Online]. Available: https://xp123.com/articles/invest-in-good-stories-and-smart-tasks/
2. B. Du Bois, S. Demeyer, J. Verelst, T. Mens, and M. Temmerman, "Does god class decomposition affect comprehensibility?" in *Proc. IASTED Conf. Softw. Eng.*, 2006, pp. 346–355.
3. N. Brousse, "The issue of monorepo and polyrepo in large enterprises," in *Companion Proc. 3rd Int. Conf. Art, Sci., Eng. Program.*, 2019, pp. 1–4, doi: 10.1145/3328433.3328435.
4. V. Gupta, E. Bjarnason, and C. Gupta, "Strategic prototyping technology adop-tion in startups: Framework, challenges, and opportunities," *IT Prof.*, vol. 24, no. 3, pp. 88–95, May/Jun. 2022, doi: 10.1109/MITP.2022.3172876.

SOFIJA HOTOMSKI is a global product manager for cloud offerings in fire Safety, 6300 Zug, Switzerland. Contact her at sofija.hotomski@siemens.com.

EDITOR: Phil Laplante, IEEE Fellow, plaplante@psu.edu

DEPARTMENT: SOFTWARE ENGINEERING

This article originally appeared in Computer vol. 58, no. 5, 2025

Citizen Development, Low-Code/No-Code Platforms, and the Evolution of Generative AI in Software Development

J. T. Sodano, EPAM Systems

Joanna F. DeFranco, The Pennsylvania State University

The demand for faster software solutions exceeds the supply of skilled software developers. More businesses will adopt citizen development frameworks and generative AI tools; however, this solution adds some challenges for project governance and security.

Many organizations are supported by software—but there is a shortage of software developers and engineers, which could cause lost revenue.[1] This is leading to empowerment in building applications using low-code/no-code (LCNC) platforms enabling faster solutions. This democratization of information technology (IT) continues to accelerate, enabling broader participation in software creation beyond traditional IT departments and software engineers.[2]

This trend has led to "citizen de-velopers," who are nontechnical users leveraging drag-and-drop tools and prebuilt components to create digital solutions for specific needs. LCNC platforms have been instrumental to this movement by providing a means for nonprofessional developers to rapidly build business applications.[3,4] Meanwhile, generative artificial intelligence (GenAI) has more recently emerged as an enabler for software development.[5] GenAI platforms, powered by large language models (LLMs), can produce or refactor code with minimal input using natural language, further reducing the barriers to entry for software innovation. However, as these tools become more intuitive and easier to use, questions arise as to whether they will introduce new risks or governance challenges, particularly in parallel with existing LCNC solutions.[5,6]

The core issue lies in determining how best to integrate the use of GenAI tools into citizen development without compromising application quality, security, and governance. LCNC platforms already address some portion of the skill gap by minimizing dependencies on complex programming languages.[7] However, the capabilities of GenAI code assistants/tools add a new layer of complexity, and the topic of how GenAI code features may complement or potentially replace traditional LCNC functionality has not yet been fully explored. In this article, we look to explore the intersecting roles of citizen development, LCNC platforms, and GenAI code systems while highlighting best practices and governance strategies that can help organizations manage the transition toward increased technology democratization if it is right for the business.

BENEFITS AND CHALLENGES OF CITIZEN DEVELOPMENT

Citizen development refers to software creation by nontechnical individuals with little to no programming skills. These are typically domain experts who lack formal training in software development and programming.[2] In many organizations, these citizen developers emerge to address gaps left by resource-constrained IT teams. Their projects often address immediate business needs such as workflow automation, data

Digital Object Identifier 10.1109/MC.2025.3547073
Date of current version: 28 April 2025

collection, and niche analytics.[3] This democratized approach can both complement and challenge the conventional enterprise IT model, where development is managed and controlled by specialized software engineers.[4]

The proliferation of citizen development has produced several tangible benefits. *Time to market* for solution innovation can be accelerated when domain experts are able to create prototypes and even entire applications more rapidly than through traditional programming scenarios, resulting in expedited digital transformation initiatives.[8] Because these citizen developers have firsthand knowledge of specific business needs, applications are often more *closely aligned with user requirements.*[9,10,11] Meanwhile, organizations often achieve *cost savings* when citizen developers reduce workloads on specialized development resources.[11,12] With increased cross-functional contribution, a heightened sense of engagement between business and IT stakeholders can *increase overall technology adoption.*[2]

Despite these advantages, some significant challenges remain. *Software quality* can vary substantially because nonexperts lack grounding in security and architecture principles.[5,10,13] *Fragmentation in governance models* also enables unmonitored "shadow IT" to grow where solutions evolve outside of sanctioned organizational oversight.[10,14] The potential for *vulnerabilities, integration issues, and application sprawl* grows absent standardized frameworks. Nevertheless, as organizations contend with market demands and IT capacity constraints, citizen development continues to increase on the basis of a growing number of tools that reduce or outright eliminate the need for coding knowledge.[9,10,11,12,13]

THE ROLE OF LCNC PLATFORMS IN CITIZEN DEVELOPMENT

LCNC platforms rely on graphical user interfaces, prebuilt modules, and configuration-driven workflows to simplify or eliminate direct source code writing.[3,4] These tools have matured significantly over time, offering features such as drag-and-drop design elements, automated database integration, and rule-based logic flows.[7] In practical terms, these features provide entry points for citizen developers by reducing the learning curve and automating a significant portion of the technical foundation typically required in traditional programming.

LCNC platforms often provide templates for common business processes and automated consistency checks. Some platforms may also integrate with advanced analytics or data visualization functions that grant users the ability to incorporate sophisticated capabilities without delving into low-level code.[10,15] While these tools may accelerate productivity, poorly governed LCNC deployments can produce redundant applications or security issues when organizations fail to coordinate efforts. Additionally, some domain experts still struggle with abstract design principles or logic flows embedded within graphical interfaces.[2] Large-scale applications introduce another level of complexity when LCNC-developed solutions must integrate with enterprise systems and adhere to the same security and performance standards as conventional software.[4] Despite these constraints, the trend toward LCNC platforms continues to grow because a structured environment where citizen developers can innovate rapidly and with minimal programming skills and fewer barriers is accessible.

THE EMERGENCE OF GENERATIVE AI CODE IN CITIZEN DEVELOPMENT

GenAI is set to redefine who can write software and how they do it, particularly in the context of citizen development. Although a low to moderate basic scripting skill level may be required, recent advances in LLMs allow AI to recommend or auto-generate entire blocks of code using natural language prompts.[5] This evolution has drawn from extensive training based on open source libraries and other code repositories, resulting in pattern-based predictions for a variety of coding tasks.[5,6]

For citizen developers, GenAI functionality can dramatically reduce complexity. If the user can state in language the desired outcome of an application, AI can propose solution logic that addresses the request.[5,15] This process supports faster prototyping and refinements, allowing people with limited coding backgrounds to iterate quickly. GenAI suggestions may also apply a framework based on best practices recognized from the AI's training corpus that decreases human errors and strengthens consistency in the final software output.[5]

Nevertheless, this approach raises a number of concerns. Even if the GenAI code syntax is correct, it *may fail to meet functional requirements* if the prompts provided are ambiguous or incomplete.[6] Within enterprise environments, questions about security and intellectual property are heightened with a *risk* that GenAI may produce code snippets that draw in part from licensed or proprietary code. Citizen developers who already struggle with verifying LCNC build solutions may find themselves even more challenged when validating machine-generated logic. In addition, the risk of unknowingly introducing malicious code or violating legal boundaries increases when users blindly accept AI outputs.[5] These issues highlight the need for further study on how GenAI can best integrate with LCNC platforms to enable organizations to benefit from faster development without compromising quality or governance.

A HYBRID MODEL INTEGRATING LCNC AND GAI TOOLS

A hybrid approach that merges the relative reliability of LCNC platforms with the versatility of GenAI models could fundamentally reshape citizen development. In this paradigm, visual workflows and structured components from LCNC systems can operate alongside real-time AI code suggestions, enabling quicker and more adaptable software development.[5,9,15] The key challenge is to integrate both in a way that respects organizational policies, safeguards security, and ensures that nontechnical contributors remain empowered rather than overwhelmed.

Comprehensive training and enablement must be part of this hybrid model. Citizen developers benefit from clear guidelines on how to craft effective prompts for AI and interpret the generated code in ways consistent with their organization's quality controls.[15] Mentorship programs that pair novices and experienced staff, or "centers of excellence," can mitigate the risks of placing too much trust in automated suggestions.[9,15] A complementary governance framework consisting of role-based access, structured reviews, and mandatory testing before production deployment can limit the potential for shadow IT scenarios.[3] In this scenario, domain experts continue to innovate while IT professionals provide oversight and ensure alignment with broader enterprise standards.

Testing and validation procedures have increased importance when combining LCNC and GenAI code capabilities. Automated tools that detect anomalies, security flaws, or accessibility issues should be run continuously as a part of the development flow.[8,16] Code reviews, assisted by separate AI modules, may be used to confirm that new logic follows best practices. In regulated industries, potential compliance enforcement tools with the ability to scan for data privacy violations could be integrated directly with LCNC platforms and GenAI engines.[5] The iterative process of automated checks followed by human validation ensures coherence across various citizen development initiatives.[16]

A WAY FORWARD

The intersection of citizen development, LCNC platforms, and GenAI represents a pivotal shift in how software is conceived, built, and governed. By extending development capabilities to a broader array of contributors, organizations can discover new paths to innovation and problem-solving during a time when competitive advantages rely on rapid digital transformation. Incorporating GenAI into these processes can further reduce barriers to entry, particularly for individuals without formal programming backgrounds.

This continuous evolution presents simultaneous challenges for project governance, application security, and the responsible use of AI-driven code suggestions. It remains uncertain how to effectively address data privacy and intellectual property concerns in this context as well as whether organizations can create standardized guidelines that balance flexibility for citizen developers with compliance with enterprise standards. To facilitate the seamless integration of GenAI into LCNC workflows, organizations should prioritize establishing best practices for oversight, implementing rigorous testing protocols, and adopting formal training programs focused on enhancing critical-thinking skills among citizen developers.

Future research must examine how best to integrate advanced AI features alongside existing LCNC functionalities. Furthermore, the organizational design implications of this hybrid development model warrant additional investigation. It is possible that new roles will emerge to serve as AI "prompt engineers," bridging the communication gap between domain experts and AI engines. Meanwhile, more sophisticated governance strategies or automated compliance mechanisms may evolve to further mitigate the challenges associated with these novel coding partnerships.

GenAI is well positioned to further the existing trends in citizen development. It has the potential to drive unprecedented innovation within enterprise software while also enhancing efficiency and empowerment for a broader and more diverse community of developers.

REFERENCES

1. K. Madding, "Developer to accelerate business efficiency," *Forbes*, Jan. 31, 2023. [Online]. Available: https://www.forbes.com/councils/forbestechcouncil/2023/01/31/the-rise-of-the-citizen-developer-to-accelerate-business-efficiency/
2. D. Hoogsteen and H. Borgman, "Empower the workforce, empower the company? Citizen development adoption," in *Proc. 55th Hawaii Int. Conf. Syst. Sci.*, 2022, pp. 4417–4726, doi: 10.24251/HICSS.2022.575.
3. S. A. A. Naqvi, M. P. Zimmer, R. Syed, and P. Drews, "Understanding the socio-technical aspects of low-code adoption for software development," in *Proc. ECIS Res. Papers, 357*. Kristiansand, Norway: Association for Information Systems, 2023. [Online]. Available: https://aisel.aisnet.org/ecis2023_rp/357
4. M. Overeem and S. Jansen, "Proposing a framework for impact analysis for low-code development platforms," in *Proc. ACM/IEEE Int. Conf. Model Driven Eng. Lang. Syst. Companion (MODELS-C)*, 2021, pp. 88–97, doi: 10.1109/MODELS-C53483.2021.00020.
5. O. Bruhin, E. Dickhaut, E. Elshan, and M. M. Li, "The rise of generative AI in low code development platforms—An analysis and future directions," in *Proc. 57th Hawaii Int. Conf. Syst. Sci.*, 2024, pp. 7780–7789, doi: 10.24251/HICSS.2023.932.
6. S. Bubeck et al., "Sparks of artificial general intelligence: Early experiments with GPT-4," 2023, *arXiv:2303.12712*.
7. C. Silva, J. Vieira, J. C. Campos, R. Couto, and A. N. Ribeiro, "Development and validation of a descriptive cognitive model for predicting usability issues in a low-code development platform," *Human Factors*, vol. 63, no. 6, pp. 1012–1032, 2021, doi: 10.1177/0018720820920429.
8. S. Rafi, M. A. Akbar, M. Sánchez-Gordón, and R. Colomo-Palacios, "DevOps practitioners' perceptions of the low-code trend," in *Proc. ACM/IEEE Int. Symp. Emp. Softw. Eng. Meas. (ESEM)*, New York, NY, USA: ACM, 2022, pp. 1–6, doi: 10.1145/3544902.3546635.
9. B. Binzer and T. J. Winkler, "Low-coders, no-coders, and citizen developers in demand: Examining knowledge, skills, and abilities through a job market analysis," in *Proc. 18th Int. Conf. Wirtschaftsinformatik*, Nuremberg, Germany: Association for Information Systems, 2023, pp. 123–130. [Online]. Available: https://aisel.aisnet.org/wi2023/17
10. N. Callinan and M. Perry, "Critical success factors for citizen development," *Open J. Appl. Sci.*, vol. 14, no. 4, pp. 1121–1149, 2024, doi: 10.4236/ojapps.2024.144073.
11. J. Kirchhoff, N. Weidmann, S. Sauer, and G. Engels, "Situational development of low-code applications in manufacturing companies," in *Proc. ACM/IEEE 25th Int. Conf. Model Driven Eng. Lang. Syst. (MODELS) Companion*, New York, NJ, USA: ACM, 2022, pp. 1–10, doi: 10.1145/3550356.3561560.
12. B. Adrian, S. Hinrichsen, and A. Nikolenko, "App development via low-code programming as part of modern industrial engineering education," in *Advances in Human Factors and Systems Interaction*, I. L. Nunes, Ed., vol. 1207, Cham, Switzerland: Springer-Verlag, 2020, pp. 45–51.
13. E. Elshan, D. Germann, E. Dickhaut, and M. Li, "Faster, cheaper, better? Analyzing how low code development platforms drive bottom-up innovation," in *Proc. ECIS Research-in-Progress Papers*. Kristiansand, Norway: Association for nformation Systems, 2023, pp. 1–10. [Online]. Available: https://aisel. aisnet.org/ecis2023_rip/82

14. M.-E. Godefroid, R. Plattfaut, and B. Niehaves, “IT outside of the IT department: Reviewing lightweight IT in times of shadow IT and IT consumerization,” in *Innovation Through Information Systems*, F. Ahlemann, R. Schütte, and S. Stieglitz, Eds., vol. 48, Cham, Switzerland: Springer-Verlag, 2021, pp. 554–571.
15. V. Berardi, V. Kaur, D. Thacker, and G. Blundell, “Towards a citizen development andragogy: Low-code platforms, design thinking, and knowledge-based dynamic capabilities,” *Int. J. Higher Educ. Manage.*, vol. 9, no. 2, pp. 1–21, 2023, doi: 10.24052/IJHEM/V09N02/ART-1.
16. V. S. Barletta, F. Cassano, A. Pagano, and A. Piccinno, “New perspectives for cyber security in software development: When end-user development meets artificial intelligence,” in *Proc. Int. Conf. Innov. Intell. Inform. Comput. Technol. (3ICT)*. Piscataway, NJ, USA: IEEE Press, 2022, pp. 531–534, doi: 10.1109/3ICT56508.2022.9990622.

J. T. SODANO is head of Digital Workplace at EPAM Systems, a leading global provider of digital engineering, software development, and consulting services headquartered in Newtown, PA 19020 USA, and a student in the Doctor of Engineering program at The Pennsylvania State University, University Park, PA 16802 USA. Contact him at jsodano@psu.edu.

JOANNA F. DEFRANCO is an associate professor of software engineering at The Pennsylvania State University, University Park, PA 16802 USA, and an associate editor in chief of *Computer*. Contact her at jfd104@psu.edu.

stay connected.

Conference Calendar

IEEE Computer Society conferences are valuable forums for learning on broad and dynamically shifting topics from within the computing profession. With over 200 conferences featuring leading experts and thought leaders, we have an event that is right for you. Questions? Contact conferences@computer.org.

NOVEMBER

1 November

- VIS (IEEE Visualization and Visual Analytics), Vienna, Austria

2 November

- FIE (IEEE Frontiers in Education Conf.), Nashville, USA
- LDAV (IEEE Symposium on Large Data Analysis and Visualization), Vienna, Austria
- QAI (IEEE Int'l Conf. on Quantum Artificial Intelligence), Naples, Italy

3 November

- ICTAI (IEEE Int'l Conf. on Tools with Artificial Intelligence), Athens, Greece
- PRDC (IEEE Pacific Rim Int'l Symposium on Dependable Computing), Seoul, Korea

6 November

- BIBE (IEEE Int'l Conf. on Bioinformatics and Bioengineering), Athens, Greece

7 November

- CSCloud (IEEE Int'l Conf. on Cyber Security and Cloud Computing), NYC, USA
- EdgeCom (IEEE Int'l Conf. on Edge Computing and Scalable Cloud), NYC, USA

10 November

- CASCON (IEEE Int'l Conf. on Collaborative Advances in Software and COmputiNg), Toronto, Canada
- ICCD (IEEE Int'l Conf. on Computer Design), Richardson, Texas, USA
- ICEBE (IEEE Int'l Conf. on E-Business Eng.), Buraydah, Saudi Arabia

11 November

- CIC (IEEE Int'l Conf. on Collaboration and Internet Computing), Pittsburgh, USA
- CogMI (IEEE Int'l Conf. on Cognitive Machine Intelligence), Pittsburgh, USA
- TPS-ISA (IEEE Int'l Conf. on Trust, Privacy and Security in Intelligent Systems, and Applications), Pittsburgh, USA

12 November

- ICDM (IEEE Int'l Conf. on Data Mining), Washington DC, USA

13 November

- ICKG (IEEE Int'l Conf. on Knowledge Graph), Limassol, Cyprus

14 November

- AI + Congress (IEEE AI + Congress), Guiyang, China

17 November

- SmartIoT (IEEE Int'l Conference on Smart Internet of Things), Sydney, Australia

21 November

- IPCCC (IEEE Int'l Performance, Computing, and Communications Conf.), Austin, USA

DECEMBER

1 December

- BDCAT (IEEE/ACM International Conf. on Big Data Computing, Applications, and Technologies), Nantes, France
- UCC (IEEE/ACM Int'l Conf. on Utility and Cloud Computing), Nantes, France

2 December

- APSEC (Asia-Pacific Software Eng. Conf.), Taipa, Macao
- RTSS (IEEE Real-Time Systems Symposium), Boston, USA

3 December

- SEC (IEEE/ACM Symposium on Edge Computing), Arlington, USA

4 December

- ICAIIHI (Int'l Conf. on Artificial Intelligence for Innovations in Healthcare Industries), Raipur, India

5 December

- ICA (IEEE Int'l Conf. on Agentic AI), Wuhan, China

8 December

- ACSAC (Annual Computer Security Applications Conf.), Honolulu, USA
- BigData (IEEE Int'l Conf. on Big Data), Macau, China

2469-7087/25 © 2025 IEEE

13 December

- iSES (IEEE Int'l Symposium on Smart Electronic Systems), Jaipur, India

14 December

- FOCS (IEEE Annual Symposium on Foundations of Computer Science), Sydney, Australia
- ICPADS (IEEE Int'l Conf. on Parallel and Distributed Systems), Hefei, China

15 December

- BIBM (IEEE Int'l Conf. on Bioinformatics and Biomedicine), Wuhan, China
- MCSoC (IEEE Int'l Symposium on Embedded Multicore/Many-core Systems-on-Chip), Singapore

17 December

- HiPC (IEEE Int'l Conf. on High Performance Computing, Data, and Analytics), Hyderabad, India

18 December

- ESAI (Int'l Conf. on Embedded Systems and Artificial Intelligence), Fez, Morocco

19 December

- ICVRV (Int'l Conf. on Virtual Reality and Visualization), Bogota, Colombia

2026

JANUARY

14 January

- ICOIN (Int'l Conf. on Information Networking), Hanoi, Vietnam

26 January

- AIxVR (IEEE Int'l Conf. on Artificial Intelligence and eXtended and Virtual Reality), Osaka, Japan

31 January

- HPCA (IEEE Int'l Symposium on High Performance Computer Architecture), Sydney, Australia

FEBRUARY

2 February

- BigComp (IEEE Int'l Conf. on Big Data and Smart Computing), Guangzhou, China

16 February

- ICNC (Int'l Conf. on Computing, Networking and Communications), Maui, USA

MARCH

6 March

- WACV (IEEE/CVF Winter Conf. on Applications of Computer Vision), Tucson, USA

16 March

- PerCom (IEEE Int'l Conf. on Pervasive Computing and Communications), Pisa, Italy

17 March

- SANER (IEEE Int'l Conf. on Software Analysis, Evolution and Reengineering), Limassol, Cyprus
- SSIAI (IEEE Southwest Symposium on Image Analysis and Interpretation), Santa Fe, USA

21 March

- VR (IEEE Conf. on Virtual Reality and 3D User Interfaces), Daegu, Korea

23 March

- SaTML (IEEE Conf. on Secure and Trustworthy Machine Learning), Munich, Germany

APRIL

10 April

- ICSE (IEEE/ACM Int'l Conf. on Software Eng.), Rio de Janeiro, Brazil

15 April

- COOL CHIPS (IEEE Symposium on Low-Power and High-Speed Chips and Systems), Tokyo, Japan

20 April

- PacificVis (IEEE Pacific Visualization Conf.), Sydney, Australia