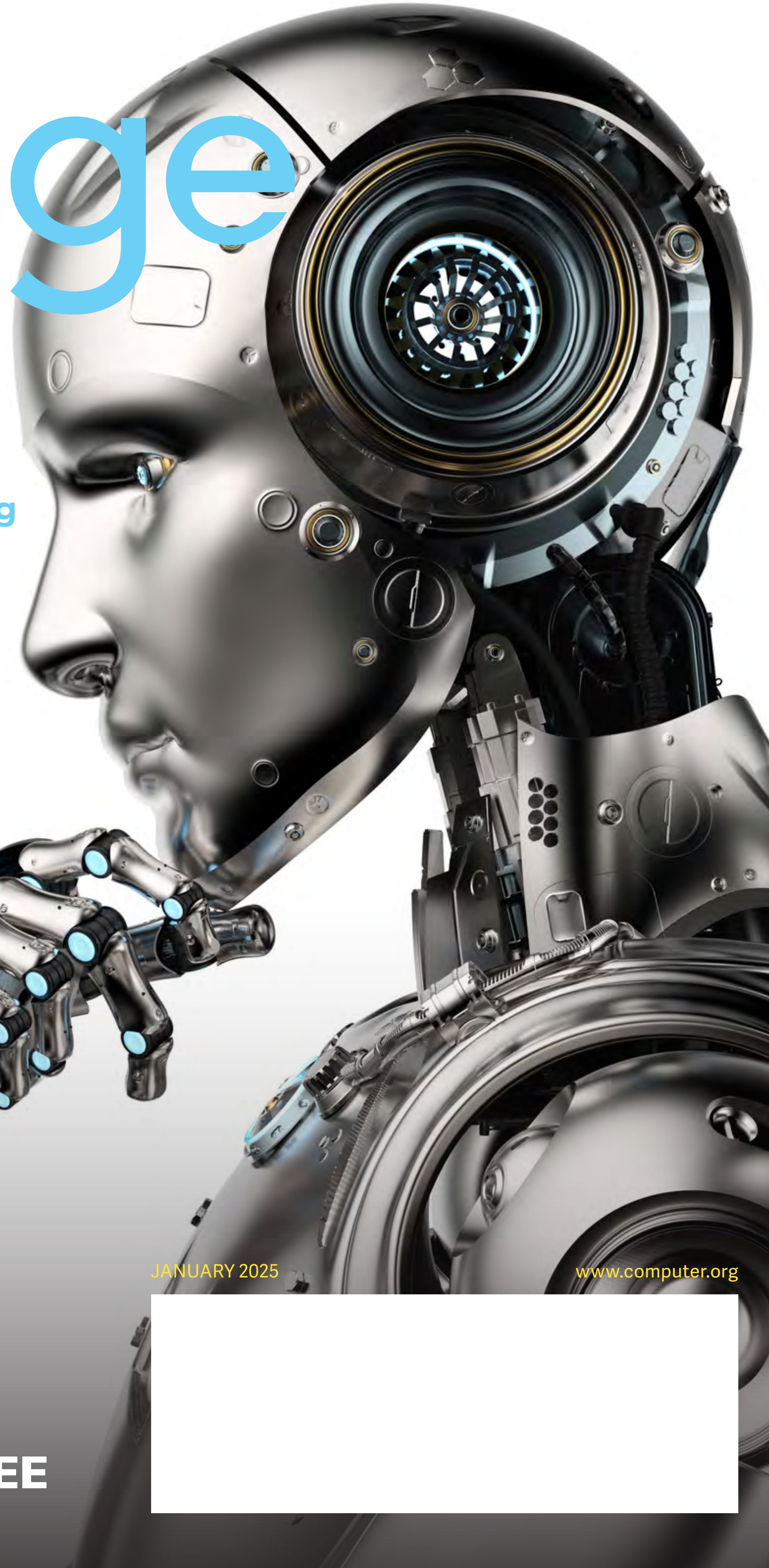
COMPUTING edge



- Machine Learning
- Distributed Computing
- Business Issues in Technology
- Blockchain

JANUARY 2025

www.computer.org

IEEE COMPUTER SOCIETY

IEEE

PUBLISH WITH THE
IEEE COMPUTER SOCIETY

Break Free. You Have Choices.

It's Author's Choice:
IEEE Computer Society provides all publishing models; open access, hybrid, and traditional options to accommodate the unique needs of all researchers.

www.computer.org/cfp

IEEE COMPUTER SOCIETY

IEEE

COMPUTING edge

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.

Reuse Rights and Reprint Permissions: Educational or personal use of this material is permitted without fee, provided such use: 1) is not made for profit; 2) includes this notice and a full citation to the original work on the first page of the copy; and 3) does not imply IEEE endorsement of any third-party products or services. Authors and their companies are permitted to post the accepted version of IEEE-copyrighted material on their own Web servers without permission, provided that the IEEE copyright notice and a full citation to the original work appear on the first screen of the posted copy. An accepted manuscript is a version which has been revised by the author to incorporate review suggestions, but not the published version with copy-editing, proofreading, and formatting added by IEEE. For more information, please go to: http://www.ieee.org/publications_standards/publications/rights/paperversionpolicy.html. Permission to reprint/republish this material for commercial, advertising, or promotional purposes or for creating new collective works for resale or redistribution must be obtained from IEEE by writing to the IEEE Intellectual Property Rights Office, 445 Hoes Lane, Piscataway, NJ 08854-4141 or pubs-permissions@ieee.org. Copyright © 2025 IEEE. All rights reserved.

Abstracting and Library Use: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy for private use of patrons, provided the per-copy fee indicated in the code at the bottom of the first page is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

Unsubscribe: If you no longer wish to receive this *ComputingEdge* mailing, please email IEEE Computer Society Customer Service at help@computer.org and type "unsubscribe *ComputingEdge*" in your subject line.

IEEE prohibits discrimination, harassment, and bullying. For more information, visit www.ieee.org/web/aboutus/whatis/policies/p9-26.html.

IEEE Computer Society Magazine Editors in Chief

Computer
Jeff Voas, *NIST*

Computing in Science & Engineering
İlkay Altintaş, *University of California, San Diego (Interim EIC)*

IEEE Annals of the History of Computing
Troy Astarte, *Swansea University*

IEEE Computer Graphics and Applications
Pak Chung Wong, *Trovares and Bill & Melinda Gates Foundation (Interim EIC)*

IEEE Intelligent Systems
San Murugesan, *Western Sydney University*

IEEE Internet Computing
Weisong Shi, *University of Delaware*

IEEE Micro
Hsien-Hsin Sean Lee, *Intel Corporation*

IEEE MultiMedia
Balakrishnan Prabhakaran, *University of Texas at Dallas*

IEEE Pervasive Computing
Fahim Kawsar, *Nokia Bell Labs and University of Glasgow*

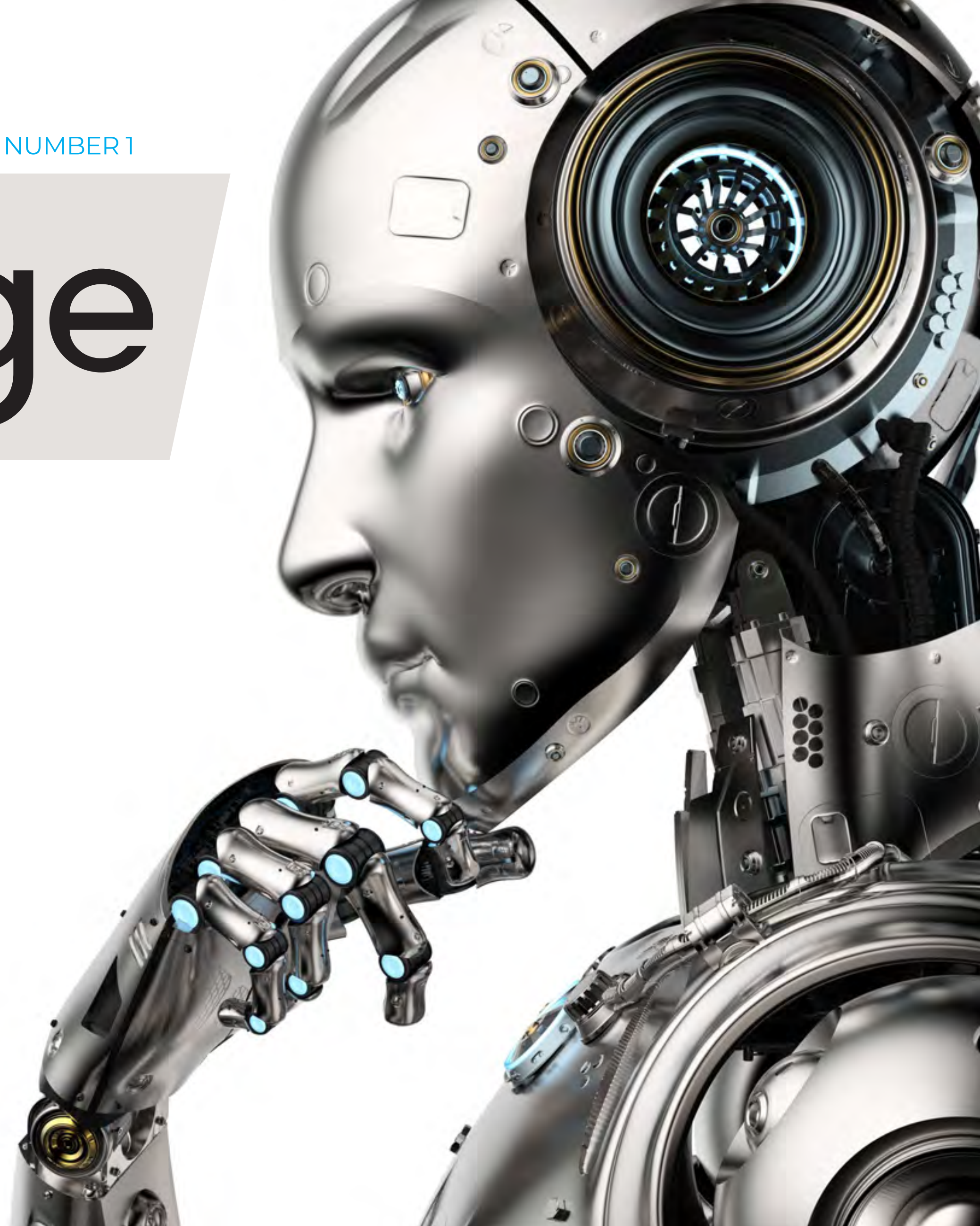
IEEE Security & Privacy
Sean Peisert, *Lawrence Berkeley National Laboratory and University of California, Davis*

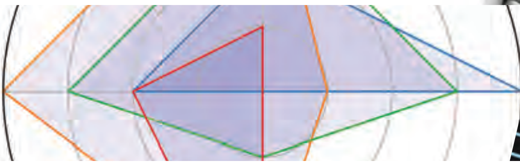
IEEE Software
Sigrid Eldh, *Ericsson, Mälardalen University, Sweden; Carleton University, Canada*

IT Professional
Charalampos Z. Patrikakis, *University of West Attica*

JANUARY 2025 · VOLUME 11 · NUMBER 1

COMPUTING edge





Machine Learning

Distributed Computing

Business Issues in Technology

Blockchain

Departments

Subscribe to *ComputingEdge* for free at
www.computer.org/computingedge

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

A Root Cause Analysis of a Self-Driving Car Dragging a Pedestrian

In this article, featured in the November 2024 issue of *Computer*, the authors use the Taxonomy for Artificial Intelligence Hazard Analysis to examine an accident on October 2, 2023, when a pedestrian was hit and dragged by one of Cruise's self-driving cars. The authors discuss the company's remote operations issues, the lack of periodic design and test protocol reviews, and Cruise's post-collision assessment deficiencies.

Computing in Science & Engineering

Exabiome: Advancing Microbial Science through Exascale Computing

The Exabiome project seeks to improve the understanding of microbiomes through the development of methods for accelerating metagenomic science using exascale computing. This April–June 2024 *Computing in Science & Engineering* article gives an overview of scientific impact of the three components of the project: metagenome assembly, protein family detection, and comparative analysis of metagenomes.

IEEE Annals of the History of Computing

Turing's Test, a Beautiful Thought Experiment

In the wake of the latest trends of AI, there has been a resurgence of claims and questions about the Turing test and its value, which are reminiscent of decades of practical "Turing" tests. This article, featured in the July–September 2024 issue of *IEEE Annals of the History of Computing*, presents a wealth of evidence, including new archival sources, and gives original answers to several open questions about Turing's 1950 paper, including its relation with early AI.

IEEE Computer Graphics and Applications

Quantum Machine Learning Playground

This article, featured in the September/October 2024 issue of *IEEE Computer Graphics and Applications*, introduces an innovative interactive visualization tool designed to demystify quantum machine learning (QML) algorithms. The authors' work is inspired by the success of classical machine learning visualization tools, such as TensorFlow Playground, and aims to bridge the gap in visualization resources specifically for the field of QML.

IEEE Intelligent Systems

Exploring Neural Networks for Musical Instrument Identification in Polyphonic Audio

This article in the September/October 2024 issue of *IEEE Intelligent Systems* introduces neural network-based methods that surpass state-of-the-art models, either by training faster or having simpler architecture, while maintaining comparable effectiveness in musical instrument identification in polyphonic music. Several approaches are presented, including two authors' proposals, i.e., spiking neural networks (SNNs) and a modular deep learning model named fully modular convolutional neural network (FMCNN).

2469-7087/25 © 2025 IEEE

IEEE Internet Computing

Requirements and Design Architecture for Digital Twin End-to-End Trustworthiness

In this July/August 2024 *IEEE Internet Computing* article, the authors discuss how digital twins (DTs) can be designed, deployed, and managed to enable end-to-end trustworthiness between applications and the physical domain. Particularly, they 1) identify the key characteristics enabling end-to-end DT trustworthiness, 2) evaluate the degree to which available DT platforms support these characteristics, 3) highlight a blueprint architecture paving the way to innovative DT platforms natively supporting end-to-end trustworthiness, and 4) show the benefits of their proposal with an industrial IoT use case.

IEEE micro

Twenty Five Years of Warehouse-Scale Computing

When Google was founded in 1998, it was already clear that successful web search would require enormous amounts of computing power and storage, and that no single computer would be able to handle this task. Consequently, its infrastructure design marked a fundamental shift toward an approach now widely embraced as warehouse-scale computing (WSC). In this article, featured in the September/October 2024 issue of *IEEE Micro*, the authors chronicle the evolution of WSC, highlighting pivotal milestones, lessons learned, and the vast opportunities that lie ahead.

IEEE MultiMedia

On Perceived AV Synchronization in 360° Multimedia

Media synchronization and, in particular, audiovisual (AV) synchronization, plays a pivotal role in multimedia systems, significantly impacting the quality of experience (QoE) perceived by users. What is intriguing is that despite the growing prevalence of multimedia content consumption in 360° environments, the issue of perceived AV synchronization remains relatively unexplored. To tackle this challenge, the authors of this July–September 2024 *IEEE MultiMedia* article present the results of a user study that assessed the influence of AV skews on QoE and the feeling of presence within 360° multimedia content.

IEEE pervasive COMPUTING
MOBILE SYSTEMS | UBIQUITOUS COMPUTING | INTERNET OF THINGS

The Future of Consumer Edge-AI Computing

In the last decade, deep learning has rapidly infiltrated the consumer end, mainly thanks to hardware acceleration across devices. However, as we look toward the future, it is evident that isolated hardware will be insufficient. Increasingly complex artificial intelligence tasks demand shared resources, cross-device collaboration, and multiple data types, all without compromising user privacy or quality of experience. To address this, the authors of this article, from the July–September 2024 issue of *IEEE Pervasive Computing*, introduce a novel paradigm centered around EdgeAI-Hub devices, designed to reorganize and optimize compute resources and data access at the consumer edge.

IEEE SECURITY & PRIVACY

Inclusive Involvement of At-Risk Users in Cybersecurity Research

This article, featured in the September/October 2024 issue of *IEEE Security & Privacy*, outlines an approach to assist cybersecurity research involving excluded at-risk users or those whose needs are overlooked. The authors bring attention to "ethics in practice" as an enabler of inclusive experimentation accounting for "human vulnerabilities" and also address "cybersecurity vulnerabilities."

IEEE Software

Generative AI to Generate Test Data Generators

Large language models (LLMs) are powerful tools for supporting developers in generating high-quality faking data. LLMs are unique systems that possibly encode 1) domain expertise, 2) testing fluency, and 3) cultural literacy. The authors of this article from the November/December 2024 issue of *IEEE Software* study the original task of using LLMs for producing fake test data. They fully implement an approach based on state-of-the-art LLM techniques for generating test data. To assess the feasibility of their approach, they curate real-world test data generation scenarios.

IT Professional

PAPR Analysis of 5G and B5G Waveforms Using Advanced PAPR Algorithms

The implementation of advanced waveforms will play an important role in enhancing the throughput, spectrum access, data rate, and capacity of the 5G and beyond 5G systems. High peak-to-average power ratio (PAPR) is a serious concern in advanced waveforms, which can drastically reduce the performance of the system. In this July/August 2024 *IT Professional* article, the authors aim to analyze PAPR algorithms when applied to advanced waveforms.

ADVERTISER INFORMATION

Advertising Coordinator

Debbie Sims
Email: dsims@computer.org
Phone: +1 714-816-2138 | Fax: +1 714-821-4010

Advertising Sales Contacts

Mid-Atlantic US, Northeast, Europe, the Middle East and Africa:
Dawn Scoda
Email: dscoda@computer.org
Phone: +1 732-772-0160
Cell: +1 732-685-6068 | Fax: +1 732-772-0164

Southwest US, California:
Mike Hughes
Email: mikehughes@computer.org
Cell: +1 805-208-5882

Central US, Northwest US, Southeast US, Asia/Pacific:
Eric Kincaid
Email: e.kincaid@computer.org
Phone: +1 214-553-8513 | Fax: +1 888-886-8599
Cell: +1 214-673-3742

Midwest US:
Dave Jones
Email: djones@computer.org
Phone: +1 708-442-5633 | Fax: +1 888-886-8599
Cell: +1 708-624-9901

Jobs Board (West Coast and Asia), Classified Line Ads

Heather Buonadies
Email: hbuonadies@computer.org
Phone: +1 623-233-6575

Jobs Board (East Coast and Europe), SE Radio Podcast

Marie Thompson
Email: marie.thompson@computer.org
Phone: +1 714-813-5094

Editor's Note

Machine Learning: Weighing the Risks and the Rewards

Machine learning (ML) is an invaluable tool that can help people become better and faster creators, workers, and engineers. Yet ML can also be harmful because it is prone to biases, inaccuracies, and cyberattacks. This issue of *ComputingEdge* grapples with the risks and rewards of using ML. The articles also discuss distributed computing, blockchain, and business issues related to technology.

Software engineers must design ML approaches that prioritize safety, reliability, and ethics. *IEEE Software*'s article, "Application of Large Language Models to Software Engineering Tasks: Opportunities, Risks, and Implications," defines large language models (LLMs) and highlights the possibilities and problems they present. In "Revisiting Edge AI: Opportunities and Challenges," from *IEEE Internet Computing*, the authors walk us through the challenges associated with incorporating AI into edge computing, as well as how to navigate the attendant design considerations.

Distributed computing is becoming more widely used. The authors of "PyCOMPSs as an Instrument for Translational Computer Science," from *Computing in Science & Engineering*, describe the PyCOMPSs project, a programming model for distributed computing, and its applications for TCS. The article, "Distributed Quantum Computing via Integrating Quantum and Classical Computing," from *Computer*, illustrates a form of distributed computing that hybridizes quantum and classical approaches.

How do key topics in business, such as the modern landscape of work and enterprise software patterns, affect employees and business leaders? "New Ways of Working Are Already Old," from *IT Professional*, argues that C-suites need to accept that remote work is here to stay and assess how to best leverage its positive attributes for innovation and business culture, even as they work to mitigate associated challenges. The author of *IEEE Software* article, "Jon Smart on Patterns and Antipatterns for Enterprise Software Success," presents a panel discussion on key topics in enterprise software, such as business agility, system entropy, and outcomes vs. outputs.

Engineers are exploring how to use blockchain efficiently and responsibly in gaming and with regards to its energy consumption. In "Integrating Blockchain Technology in Online Gaming Ecosystems," from *Computer*, the author insists that research and collaboration between blockchain and game developers are essential to addressing scalability, regulatory compliance, and system integration. *Computer* article "Blockchain's Carbon and Environmental Footprints" analyzes blockchains' energy consumption and the resulting environmental impacts.

2469-7087/25 © 2025 IEEE

EDITOR: Ipek Ozkaya, Carnegie Mellon Software Engineering Institute,
ipek.ozkaya@computer.org

This article originally appeared in
IEEE Software
vol. 40, no. 3, 2023

DEPARTMENT: FROM THE EDITOR

Application of Large Language Models to Software Engineering Tasks: Opportunities, Risks, and Implications

Ipek Ozkaya

Has the day we all have been waiting for really arrived? Have advances in deep learning and machine learning (ML) finally reached a turning point and have started to produce "accurate enough" assistants to help us in a variety of tasks, including software development? Are large language models (LLM) going to turn us all into better writers, artists, translators, programmers, health-care workers, not to mention software engineers? Or are we at a risky turning point where we will not be able to separate artificial intelligence (AI)-generated content from user-created ones, drowning in misinformation and perfect sounding yet fake and incorrect information and AI-generated faulty programs?

Recently released LLMs, such as Generative Pretrained Transformer (GTP) 4 used in ChatGPT by OpenAI and BERT used in Bard by Google, disrupt the search engine model that we have been used to. Use of these models shifts the end-user computer interaction from "here are a list of places to look at to potentially find an answer to your question" to "here is a suggested answer to your questions with well-constructed syntax, what is your next question based on this?"

Without a doubt, LLMs have use cases in assisting software engineering tasks as well, including code generation models trained in programming languages, such as CoPilot by GitHub. The reaction of the software engineering community to the accelerated advances that LLMs have been enjoying since 2022 has been varied, ranging from considering capabilities offered by these models as "snake oil"[1] to "end of programming and computer science education as we know it."[2] In this article, after a brief overview of LLMs, I will focus on the opportunities LLMs open up for software development and implications of incorporating LLMs into systems as well as assisting with software development tasks.

WHAT ARE LMMs?

An LLM is a deep neural network model which has been trained on large amounts of data, such as books, code, articles, and websites, to learn the underlying patterns and relationships in the language that it was trained for. By doing so, the model is able to generate coherent content such as grammatically correct sentences and paragraphs that mimic human language or syntactically correct code snippets. LLMs have applications in a variety of tasks, including language translation, summarization, and question answering and have potential in many fields as long as the data that the models have been trained on provide the appropriate input. While the content generated by LLMs are often grammatically correct, they may not always be semantically correct. The probabilistic and randomized selection of the "next token" in constructing the outputs on one hand gives the end user the impressions of correctness and style, on the other hand may result in mistakes.[3]

While the recently released versions of LLMs, ChatGPT driving the pack, have made significant improvements, there are several areas of caution around their generation and use:

- *Data quality and bias concerns:* LLMs require enormous amounts of training data to learn language patterns and their outputs are highly dependent on the data that they are trained on.

Digital Object Identifier 10.1109/MS.2023.3248401
Date of current version: 18 April 2023

2469-7087/25 © 2025 IEEE

Any of the issues that exist in the training data, such as biases and mistakes, will be amplified by LLMs, potentially resulting in models that exhibit discriminatory behavior, such as making prejudiced recommendations. This means that the quality and representativeness of the training data can significantly impact the model's performance and generalizability, mistakes can propagate. For example, language models that are used to recommend code patterns have been found to carry security flaws forward.[4] This creates risks in not only generating buggy code, but also perpetuating immature implementation practices in developers.

- *Privacy and content ownership concerns:* LLMs are generated using content developed by others which both may contain private information as well as content creators' unique creativity characteristics. Training on such data using patterns in recommended output creates plagiarism concerns. Some content is boiler plate and the ability to generate output in correct and understandable ways creates opportunities for improved efficiency. But content, including code, where individual contributions matter becomes difficult to differentiate. In the long run, increasing popularity of language models will likely create boundaries around data sharing and open source software and open science. Techniques to indicate ownership or even preventing certain data to be used to train such models will likely emerge. However, such techniques and attributes to complement LLMS are yet to come.
- *Environmental concerns:* The vast amounts of computing power required in training deep learning models has been increasingly a concern related to their impact on carbon footprint. Research in different training techniques, algorithmic efficiencies, and varying allocation of computing resources during training will likely increase. In addition, improved data collection and storage techniques are anticipated to eventually reduce the impact of LLMs on the environment, but development of such techniques are still in their early phases.[5]
- *Explainability and unintended consequence concerns:* Explainability of deep learning and ML models is a general concern in AI, including but not limited to LLMs. Users seek to understand the reasoning behind the recommendations, especially if such models are to be used in safety or business critical settings. Dependence on the quality of the data and inability to trace the recommendations to the source increase trust concerns.[6] In addition, since the sequences are generated using a randomized probabilistic approach, explainability of correctness of the recommendations create added challenges. Explainability as well as responsible AI practices are critical since such models can easily be used to spread misinformation.

The application programming interfaces (API) of GPT and BERT are now also available to other developers. This contributes to both accelerating the use and improvements on LLMs as well as increasing the number of opportunities of their misuse. OpenAI researchers are open about their lessons learned and have no choice but rely on software engineering best practices. They recommend policy enforcement as a mechanism to enforce avoiding misuses.[7] Applications which help detect text written by such models have been quick to come, such as GPTZero written for educators to detect such text, and ironically it uses ChatGPT in doing so.[8] It is safe to say LLMs have attracted a fair share of confusion, criticism, and excitement all at the same time.

APPLICATIONS IN SOFTWARE ENGINEERING

Research agendas developed recently had already shined the light on the future of software engineering to be an AI-augmented development lifecycle where both software engineering and AI assistants share roles from copilot to student, expert, and supervisor.[9] In the National Agenda for Software Engineering, my colleagues and I had suggested that developers will need to guide and consequently improve the AI assistants. AI assistants will also take on a supervisory role by providing real-time feedback and, in time, demonstrating repeated mistakes to developers. On a developer team, there will always be some developers who you trust more than others (perhaps due to experience, skill sets, or demonstrated performance). The AI-assisted development workflows will trigger the need to think of AI "partners" in the same way.[9]

While with caution, software engineers need to think about LLMs as partners and focus on where their optimal application can be. There are quite a number of software engineering tasks which can effectively benefit from using LLMs. Indulge me for a moment to assume that we solved the trust and unethical use issues as I enumerate potential use cases where LLMs can create strides of advances in improved productivity of software engineering tasks, and where the risks can still be manageable.

- *Specification generation:* Quite a number of requirements can be common across applications, yet oftentimes requirements are also incomplete. LLMs can assist in generating more complete specifications significantly quicker.
- *Just in time developer feedback:* Applications of LLMs in software development has been received with much skepticism, rightly so at the time being. While the code generated by current AI assistants, such as Copilot, have been found to carry more security issues,[4] in time this will change. AI-based and other approaches which give developers syntactic corrections and suggestions have been around a while. LLMs carry the promise of going the extra mile and recommending not just corrections, but next steps.
- *Improved testing:* Generating unit tests is one of the tasks where developers shortcut the most. Ability to generate test cases at ease would increase overall test effectiveness and coverage, and consequently system quality.
- *Documentation:* Ranging from contracting language to regulatory requirements, there are many applications of LLMs to software development documentation.
- *Language translation:* Legacy software and brownfield development is the norm of system development today, and many organizations need to go through language translation efforts when they need to modernize their systems. This process is often manual and error prone, while some tools do exist to support developers. While will not work at scale, portions of code can potentially be translated to other programming languages using LLMs. Rewriting a system in an other programming language is not just a language translation exercise, it is mostly also a re-architecting exercise; however, ability to rewrite selected portions at ease would be a welcomed capability.

LLMs will also require software engineers to become more savvy in how they incorporate them into systems as elements. Example areas include the following:

- *LLMs as functional components:* LLMs will definitely change some of the ways capabilities are bundled and delivered as well, where pretrained models become parts of systems or parts of external systems. APIs to LLMs will drive different system composition scenarios and will be available as services.
- *Operations informing development:* Data is the first-class citizen in LLM tools. Operational data will need to be more timely fed back to both the development process, e.g., areas where users make most mistakes, as well as functionality development, e.g., inform functionality that users do not use to be deprecated.

These examples focus on existing software engineering tasks that can be done better or faster because such models exist. There are also, however, task flows that will change, and new activities will likely emerge while time spent on others get reduced. An AI-augment software development lifecycle will likely have different task flows, efficiencies, and roadblocks than the current development lifecycles of agile and iterative development workflows. For example, rather than thinking about steps of development as requirements, design, implementation, test, and deploy, LLMs can enable bundling these tasks together. This would change the number of hand-offs and where they happen, shifting task dependencies within the software development lifecycle.

GOING FORWARD

All the areas of cautions and risks related to LLMs are areas where we need new research and innovations. These need to be targeted at improving correctness of LLM recommendations, improving their generalizability, as well as improving the ethical implications of data use and content creation.

We are likely to see most advances in generalizability of models, development of integrated development environments with new paradigms, and reliable data collection and use techniques in the near future. Curricula development and education of the next generation of computer scientists and software engineers cannot stay blind to the implications of such developments in generative AI either.

Generalizability of Models

Currently, LLMs work by pretraining on a large corpus of content followed by fine-tuning on a specific task. What this implies is that the architecture of the model is task independent; however, its application for specific tasks requires further fine-tuning with significantly large numbers of examples. Generalizability of these models to applications where data are sparse, few-shot settings, is already a focus area by researchers.[10]

New Development Environments

If we are convinced by the argument that some tasks can be accelerated and improved in correctness by AI assistants including LLMs, that also implies that the current integrated development tools will need to incorporate these assistants. When assistants are integrated in, then development becomes a more interactive process with the tool environment. Software engineering bots are already pushing the envelope of the development environments in the direction of incorporating developer assistants.[11]

Data as a Unit of Computation

The most critical input which drives this next generation of AI innovations is not only the algorithms, but also data. Not only will a significant portion of computer science and software engineering talent shift to data science and data engineer careers, but also, we will need more tool-supported innovations in data collection, data quality assessment, and data ownership rights management. This is an area with huge gaps that requires skill sets that span computer science, policy, engineering, as well as deep knowledge in security, privacy, and ethics.

Computer Science and Software Engineering Education

The biggest implications of LLMs are in how we teach programming languages and system design. LLMs are likely to take already existing platforms such as StackOverflow and Reddit, which have become indispensable resources for developers, to a new level of reduced barrier of entry. Computer science and software engineering programs need to start a shift in their curricula today. Software engineering and computer science education has already missed the boat by continuing to focus on teaching green field development while today the reality of system development is brownfield. Students are not adequately exposed to theories and techniques to support system development by composition, legacy evolution, and using heterogeneous platforms and programming languages in concert. We teach students hello world development, while we should be teaching them how to read millions of lines of code, triage and fix bugs that they have not contributed to and understand the structure and behavior of the software rather than the single class or story card they are responsible for. With LLMs and their sister AI-driven apps assisting developers, we need to be teaching next-generation software engineers when to trust, how to create evidence to trust, how to do trust assessment rapidly and correctly, and how to improve such assistants. We need to teach them how to evolve systems to incorporate such components, and we need to teach them to treat data as code. We need to make ethics courses mandatory every year of the curriculum. The list goes on.

COMPUTER SCIENCE AND SOFTWARE ENGINEERING PROGRAMS NEED TO START A SHIFT IN THEIR CURRICULA TODAY.

After the two winters of AI, generally attributed to late 1970s and early 1990s, we have entered not only a period of AI blossoms, but also exponential growth in funding, in use, and in scare from AI. Advances in LLMs without a doubt are huge contributors to this growth. What will determine if the next phase includes innovations beyond our imagination or another AI winter is largely dependent on not our ability to continue technical innovations, but on our ability to practice software engineering and computer science through the highest level of ethics and responsible practices. We need to be bold in experimenting with the potential of LLMs in improving software development, and we need to be cautious and not forget fundamentals of engineering ethics and rigor.

REFERENCES

1. S. Shankland. "Computing guru criticizes ChatGPT AI tech for making things up." CNET. Accessed: Feb. 2023. [Online]. Available: https://www.cnet.com/tech/computing/computing-guru-criticizes-chatgpt-ai-tech-for-making-things-up/
2. M. Welsh, "The end of programming," *Commun. ACM*, vol. 66, no. 1, pp. 34–35, Jan. 2023, doi: 10.1145/3570220.
3. S. Wolfram. "What is ChatGPT doing … and why does it work?" Stephen Wolfram. Accessed: Feb. 2023. [Online]. Available: https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/
4. N. Perry, M. Srivastava, D. Kumar, and D. Boneh, "Do users write more insecure code with AI assistants?" 2022, *arXiv:2211.03622*.
5. D. A. Patterson et al., "The carbon footprint of machine learning training will plateau, then shrink," *Computer*, vol. 55, no. 7, pp. 18–28, Jul. 2022, doi: 10.1109/MC.2022.3148714.
6. C. Tantithamthavorn, J. Cito, H. Hemati, and S. Chandra, "Explainable AI for SE: Experts' interviews, challenges, and future directions," *IEEE Softw.*, vol. 40, no. 4, 2023.
7. M. Brundage et al., "Lessons learned on language model safety and misuse." OpenAI. Accessed: Feb. 2023. [Online]. Available: https://openai.com/blog/language-model-safety-and-misuse/
8. GPTZero. Accessed: Feb. 2023. [Online] Available: https://gptzero.me/faq
9. A. Carleton et al., "Architecting the future of software engineering: A national agenda for software engineering research and development," Softw. Eng. Inst., Pittsburgh, PA, USA, AD1152714, 2021.
10. T. B. Brown et al., "Language models are few-shot learners," 2020, *arXiv:2005.14165*.
11. I. Ozkaya, "A paradigm shift in automating software engineering tasks: Bots," *IEEE Softw.*, vol. 39, no. 5, pp. 4–8, Sep./Oct. 2022, doi: 10.1109/MS.2022.3167801.

IEEE COMPUTER SOCIETY
Call for Papers

Build your authority in the industry with exposure to a global network of 350K+ computing professionals.

GET PUBLISHED
www.computer.org/cfp

IEEE COMPUTER SOCIETY
IEEE

IEEE COMPUTER SOCIETY

www.computer.org

PURPOSE: Engaging professionals from all areas of computing, the IEEE Computer Society sets the standard for education and engagement that fuels global technological advancement. Through conferences, publications, and programs, IEEE CS empowers, guides, and shapes the future of its members, and the greater industry, enabling new opportunities to better serve our world.

OMBUDSMAN: Contact ombudsman@computer.org.

CHAPTERS: Regular and student chapters worldwide provide the opportunity to interact with colleagues, hear technical experts, and serve the local professional community.

PUBLICATIONS AND ACTIVITIES

***Computer*:** The flagship publication of the IEEE Computer Society, *Computer*, publishes peer-reviewed technical content that covers all aspects of computer science, computer engineering, technology, and applications.

Periodicals: The society publishes 12 magazines, 18 journals

Conference Proceedings & Books: Conference Publishing Services publishes more than 275 titles every year.

Standards Working Groups: More than 150 groups produce IEEE standards used throughout the world.

Technical Communities: TCs provide professional interaction in more than 30 technical areas and directly influence computer engineering conferences and publications.

Conferences/Education: The society holds more than 215 conferences each year and sponsors many educational activities, including computing science accreditation.

Certifications: The society offers three software developer credentials.

AVAILABLE INFORMATION

To check membership status, report an address change, or obtain information, contact help@computer.org.

IEEE COMPUTER SOCIETY OFFICES

WASHINGTON, D.C.:
2001 L St., Ste. 700,
Washington, D.C. 20036-4928
Phone: +1 202 371 0101
Fax: +1 202 728 9614
Email: help@computer.org

LOS ALAMITOS:
10662 Los Vaqueros Cir.,
Los Alamitos, CA 90720
Phone: +1 714 821 8380
Email: help@computer.org

IEEE CS EXECUTIVE STAFF

Executive Director: Melissa Russell
Director, Governance & Associate Executive Director: Anne Marie Kelly
Director, Conference Operations: Silvia Ceballos
Director, Information Technology & Services: Sumit Kacker
Director, Marketing & Sales: Michelle Tubb
Director, Membership Development: Eric Berkowitz
Director, Periodicals & Special Projects: Robin Baldwin

IEEE

revised 29 October 2024

IEEE CS EXECUTIVE COMMITTEE

President: Jyotika Athavale
President-Elect: Hironori Washizaki
Past President: Nita Patel
First VP: Grace A. Lewis
Second VP: Nils Aschenbruck
Secretary: Mrinal Karvir
Treasurer: Darren Galpin
VP, Member & Geographic Activities: Kwabena Boateng
VP, Professional & Educational Activities: Cyril Onwubiko
VP, Publications: Jaideep Vaidya
VP, Standards Activities: Edward Au
VP, Technical & Conference Activities: Terry Benzel
2023–2024 IEEE Division VIII Director: Leila De Floriani
2024–2025 IEEE Division V Director: Christina M. Schober
2024 IEEE Division V Director-Elect: Thomas M. Conte

IEEE CS BOARD OF GOVERNORS

Term Expiring 2024:
Saurabh Bagchi, Charles (Chuck) Hansen, Carlos E. Jimenez-Gomez, Daniel S. Katz, Shixia Liu, Cyril Onwubiko

Term Expiring 2025:
İlkay Altintaş, Mike Hinchey, Joaquim Jorge, Rick Kazman, Carolyn McGregor, Andrew Seely

Term Expiring 2026:
Megha Ben, Terry Benzel, Mrinal Karvir, Andreas Reinhardt, Deborah Silver, Yoshiko Yasuda

IEEE EXECUTIVE STAFF

Executive Director and COO: Sophia Muirhead
Interim General Counsel and Chief Compliance Officer: Ahsaki Benion
Chief Human Resources Officer: Cheri N. Collins Wideman
Managing Director, IEEE-USA: Russell Harrison
Chief Marketing Officer: Karen L. Hawkins
Managing Director, Publications: Steven Heffner
Staff Executive, Corporate Activities: Donna Hourican
Managing Director, Member and Geographic Activities: Cecelia Jankowski
Chief of Staff to the Executive Director: Kelly Lorne
Managing Director, Educational Activities: Jamie Moesch
IEEE Standards Association Managing Director: Alpesh Shah
Chief Financial Officer: Thomas Siegert
Chief Information Digital Officer: Jeff Strohschein
Managing Director, Conferences, Events, and Experiences: Marie Hunter
Managing Director, Technical Activities: Mojdeh Bahar

IEEE OFFICERS

President & CEO: Thomas M. Coughlin
President-Elect: Kathleen Kramer
Past President: Saifur Rahman
Director & Secretary: Forrest D. Wright
Director & Treasurer: Gerardo Barbosa
Director & VP, Publication Services & Products: Sergio Benedetto
Director & VP, Educational Activities: Rabab Kreidieh Ward
Director & VP, Membership and Geographic Activities: Deepak Mathur
Director & President, Standards Association: James E. Matthews III
Director & VP, Technical Activities: Manfred J. Schindler
Director & President, IEEE-USA: Keith A. Moore

This article originally appeared in
IEEE Internet Computing
vol. 28, no. 4, 2024

Revisiting Edge AI: Opportunities and Challenges

Tobias Meuser, *Technical University of Darmstadt, 64283, Darmstadt, Germany*

Lauri Lovén, *University of Oulu, 90014 Oulu, Finland*

Monowar Bhuyan, *Umeå University, 90187 Umeå, Sweden*

Shishir G. Patil, *UC Berkeley, California, Berkeley, CA, 94709, USA*

Schahram Dustdar, *Vienna University of Technology, Vienna 1040, Austria*

Atakan Aral, *Umeå University, 90187 Umeå, Sweden and University of Vienna, Vienna 1090, Austria*

Suzan Bayhan, *University of Twente, 7500 AE, Enschede, The Netherlands*

Christian Becker, *University of Stuttgart, 70569, Stuttgart, Germany*

Eyal de Lara, *University of Toronto, Toronto, ON, M5S 1A1, Canada*

Aaron Yi Ding, *TU Delft, 2600 AA, Delft, The Netherlands*

Janick Edinger, *University of Hamburg, 22527, Hamburg, Germany*

James Gross, *Royal Institute of Technology Stockholm (KTH), 100 44 Stockholm, Sweden*

Nitinder Mohan, *Technical University of Munich, 80333, Munich, Germany*

Andy D. Pimentel, *University of Amsterdam, 1098 GH, Amsterdam, The Netherlands*

Etienne Rivière, *UCLouvain, B-1348, Louvain-la-Neuve, Belgium*

Henning Schulzrinne, *Columbia University, New York, NY, 10027, USA*

Pieter Simoens, *Ghent University-imec, B-9052, Gent, Belgium*

Gürkan Solmaz, *NEC Laboratories Europe, 69115, Heidelberg, Germany*

Michael Welzl, *University of Oslo, 0313, Oslo, Norway*

Edge artificial intelligence (AI) is an innovative computing paradigm that aims to shift the training and inference of machine learning models to the edge of the network. This paradigm offers the opportunity to significantly impact our everyday lives with new services such as autonomous driving and ubiquitous personalized health care. Nevertheless, bringing intelligence to the edge involves several major challenges, which include the need to constrain model architecture designs, the secure distribution and execution of the trained models, and the substantial network load required to distribute the models and data collected for training. In this article, we highlight key aspects in the development of edge AI in the past and connect them to current challenges. This article aims to identify research opportunities for edge AI, relevant to bring together the research in the fields of artificial intelligence and edge computing.

1089-7801 © 2024 IEEE
Digital Object Identifier 10.1109/MIC.2024.3383758
Date of current version 2 August 2024.

Edge computing is a significant paradigm shift that is reshaping the internet and applications landscapes by bringing data processing closer to the source of the data. This strategic evolution has the potential to enhance efficiency, responsiveness, and the respect of privacy.[2,3,4,21] Starting from mostly cloud-based solutions, more and more applications are now pushed along the computing continuum, closer and closer to edge devices. While there have been several definitions of the latter in the past, ranging from user end devices up to small, localized data centers, the general properties of edge devices are similar: their closeness to the user and the locality of processed data.[2] While the popularity of edge solutions increased in the recent years, the deployment of edge solutions is still relatively slow compared to the growth of the cloud market. This can be attributed to the high cost of building and managing a distributed infrastructure, but also to the relative complexity of building applications for the edge compared to building them only for the cloud.

The emergence of artificial intelligence (AI) and its significant demand for training data has made the use of edge devices for training and inference a clear subsequent development.[a] The requirements of machine learning (ML) applications for vast amounts of data make, indeed, the training and inference using that data at the edge an efficient and reasonable option in comparison to a cloud-centric approach. In addition, training and inference of ML models close to or at the edge come with significant advantages for end users, including better respect for data privacy and faster response times. However, the combination of artificial intelligence with edge computing also opens further challenges, especially due to the resource constraints and availability of those edge devices. These limitations are even more evident when comparing edge devices with the robust and omnipresent cloud infrastructure. Yet, applications like autonomous driving, which demand low-latency responses as well as processing of very high-dimensional data at very high rates, vividly illustrate the necessity of edge intelligence. In such safety-critical applications, even milliseconds matter, making it essential to have access to data sources and model decisions with minimal delay. Similarly, bringing learning and inference to the edge will enable new, innovative, and useful applications such as robotics, immersive multi-user applications (augmented reality), and smart health care, revolutionizing our way of living.

While exploring the synergy of AI and edge computing, it is crucial to address the unique challenges the integration of edge computing and intelligence presents. Despite its potential, edge intelligence can be influenced by resource constraints, notably in computing and storage resources, which are in significant contrast to the capabilities of traditional cloud infrastructures. Due to these limitations, protecting data and ensuring fast response times remain significant challenges, in which today's edge computing solutions are still outperformed by pure cloud-based computing on many occasions.[21] Edge infrastructure is usually deployed in physically accessible places and cannot benefit from the perimeter-based protection measures used in cloud computing. To make the edge a real augmentation for current cloud-only solutions, future research is necessary, focusing on the security, availability, and efficiency of edge intelligence. This articles not only reviews the decade-long journey to edge AI but also critically examines the viewpoints of various stakeholders and outlines the pressing challenges and exciting future research directions in this field.

THE DECADE-LONG JOURNEY TO EDGE AI

Edge intelligence emerged as an evolution of the edge computing paradigm, whose roots are traceable to the 2000s, primarily driven by the limitations of cloud computing in handling the burgeoning data generated by local devices, e.g., the Internet of Things (IoT). Edge computing decentralizes data processing, pushing it closer to data sources at the network's edge. This proximity reduces the distance data must travel, thereby decreasing latency and conserving bandwidth. Furthermore, edge computing alleviates the data load on central servers and enhances privacy by processing sensitive data locally.[4] Edge and cloud computing can complement each other and form the so-called continuum, with edge computing addressing immediate, localized processing needs while cloud computing remains essential for large-scale data storage and extensive computational tasks.

Advent of Edge AI

Edge intelligence represents a further paradigm shift from edge computing, integrating AI to enhance the processing capabilities at the edge of the network. This integration further reduces latency and alleviates the bandwidth demand on central servers, while also providing additional benefits, such as enhanced privacy

[a] While there is debate about the differences between edge AI and edge intelligence, we use the terms edge AI and edge intelligence interchangeably in this article.

due to distributed approaches for ML like federated learning[9] and improved resilience due to local autonomy and decentralized control.[5] Edge intelligence has applications in various domains, including smart cities, health care, autonomous driving, and industrial automation, where low latency and local data processing are critical. This trend is further augmented by the increasing prevalence of 5G networks and the promises of future 6G networks, which offer the high-speed connectivity necessary for edge intelligence applications.[7] Figure 1 presents an illustration of the shift from a centralized, cloud-based use of AI for training and inference, to edge AI solutions in two representative use cases: autonomous driving and connected health solutions.

Edge AI Today

The state-of-the-art in edge intelligence can be divided into two main subfields: *AI on edge*, focusing on AI methods suitable for the decentralized, heterogeneous, and opportunistic edge environment, and *AI for edge*, focusing on the use of those methods for the benefit of the computing continuum.[6]

AI on edge has been propelled by advances in ML algorithms, particularly in deep learning, and their optimization for execution on constrained devices. The development of lightweight neural networks and techniques like model pruning and quantization are crucial in enabling complex AI models to run efficiently at the edge. In Figure 1, AI on edge allows model training and inference directly at the edge, either in a collaborative form through direct interaction between edge devices or using local edge servers close to these devices.

A notable trend is the emergence of distributed ML techniques for training and inference of AI models across multiple edge devices while preserving data privacy. For example, *federated learning* enables collaborative model training without the need to centralize data, aligning with the distributed nature of edge computing and addressing growing concerns around data security and privacy in AI.[9] To perform inference of large AI models at the edge without compressing them via pruning or quantization, these models can be split into several submodels. This allows for their distributed and collaborative execution on multiple, possibly heterogeneous, edge devices.[17,19] Alternatively, one may explore adaptive computation techniques where the inference cost is a function of the complexity of the data.[20] Finally, *hierarchical inference*[22] has been proposed where the interplay between larger and smaller neural network structures is leveraged toward accuracy, energy efficiency, and latency in edge-based inference scenarios.[23]

AI for edge, on the other hand, has seen significant advancements in integrating artificial intelligence with

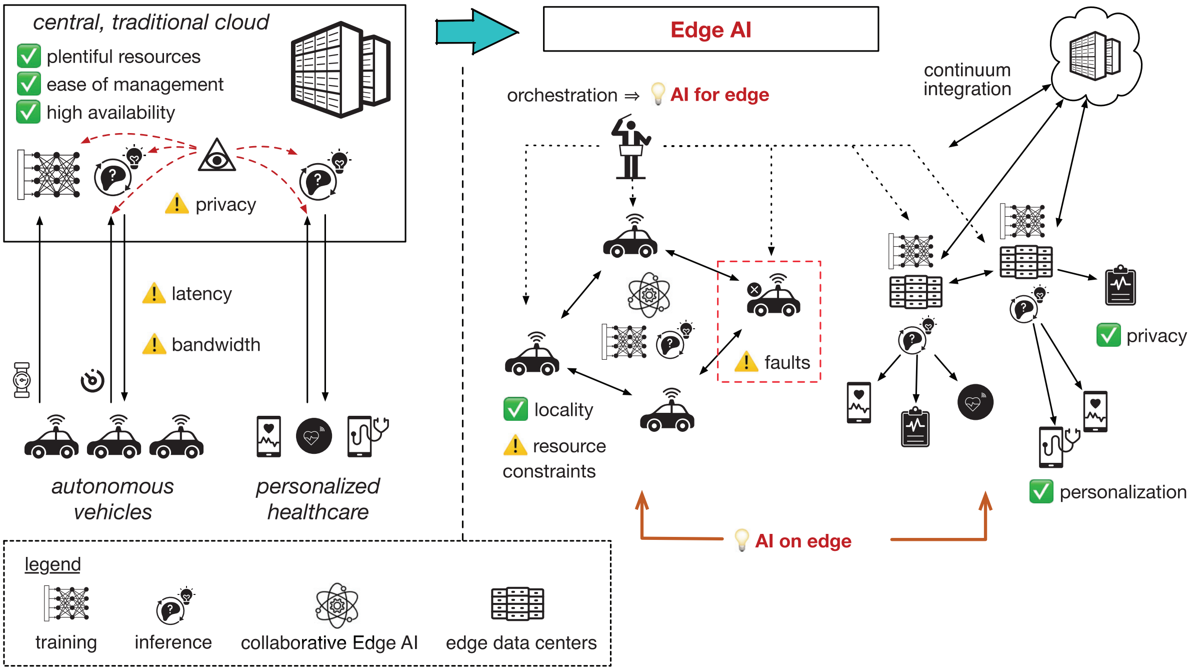


FIGURE 1. An illustration of the shift from centralized AI in the cloud (left) and Edge AI (right), and the associated challenges and opportunities, for two representative target applications: autonomous vehicles and personalized health care.

edge computing architectures, enhancing the capability of edge devices to perform sophisticated data processing and decision-making tasks, and paving the way for the intelligent orchestration of resources in the computing continuum,[10] as illustrated in Figure 1. Indeed, in addition to technological advancements, the current landscape of edge intelligence is shaped by an increasing focus on energy efficiency and sustainability.[8] Researchers and practitioners are actively exploring methods to reduce the energy footprint of edge AI systems. This is crucial for their widespread deployment, particularly in environments where power availability is a constraint. Necessary progress includes, for example, the development of energy-aware algorithms and hardware optimizations.

Besides the characterization of AI on edge or AI for edge, we observe differences in provider models. Many applications of edge computing are extensions of multitier architectures that shift the processing along the continuum between sensors and actuators, coordination of the application domain, e.g., a production floor, and cloud services. Edge computing offers the opportunity to conquer communication load and latency requirements with the placement of processing along this continuum. As such, we see edge AI as a phenomenon in industrial applications.

WHO SHOULD CARE?

While the importance of edge computing and edge AI increased in the last decade with the introduction of increasingly challenging and data-driven applications like smart cities and industrial automation, different stakeholders have different perspectives on these paradigms and associated technologies. In the following, we introduce the perspectives of four stakeholders: the needs of *society* and *industry* are shaped into solutions by *developers*. These solutions are then subject to policies and regulations set by *governments*. Understanding the individual perspectives of these stakeholders is pivotal for shaping future research directions and enabling the sound development of Edge AI.

Figure 2 provides an overview of the interest of the societal, governmental, industrial, and developer perspectives for the different challenges of Edge AI. The plot should be interpreted as a general tendency.

Societal Perspective (Everyday Life of People)

Societally, the interest in Edge AI centers on its *practical applications* rather than on the underlying technological innovations. People are likely to appreciate the use of Edge AI in areas such as autonomous vehicles and

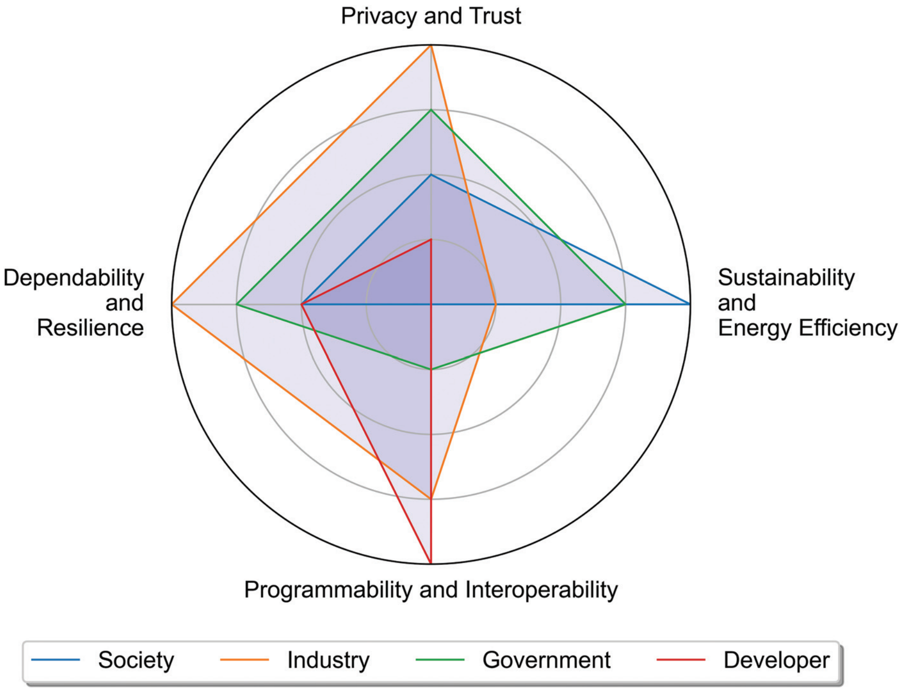


FIGURE 2. Demands of the societal, governmental, industrial, and developer perspectives. The darkness of the symbol illustrates the importance of each demand for the respective stakeholder's perspective (darker color means higher demand).

smart homes. Although the average users, particularly those without a technical background, may not notice the latency differences between cloud-based and edge-based execution, the accessibility of applications and their impact on daily life will be much more significant, as it will enable richer interactions and more complex applications. Especially in Europe, *privacy* is an important aspect, which is closely linked with edge computing and edge AI.

Industry Perspective

We can distinguish perspectives for two categories of industrial players: *consumers* of edge AI and *providers* of edge AI.

For consumers of edge AI, the question of *reliability and guarantees* is a major factor in deciding the future use of edge intelligence. The multitude of cloud and edge providers makes it hard to have confidence in the reliable operation of multiple systems and services. In addition, the ability to attribute system failures to specific components or providers diminishes, thereby limiting the potential for liability for such failures. Although this challenge can be mitigated by using combined cloud/edge providers, such as AWS Wavelength, using a single provider may substantially impact some of the benefits of edge computing, particularly in terms of system robustness and data protection. *Data protection* is a factor that, similarly to the societal perspective, is critical from an industrial viewpoint. This includes the protection of data while being processed at the edge, ensuring the *trustworthiness* of the edge network provider, and the protection of intellectual property, i.e., of the developed edge applications and trained AI models.[1]

For providers of edge AI, the question of *business cases* is pivotal for the success of edge AI. In the past, there already has been a transition from voice providers to data providers in telecommunication, who can now again transition, this time to computation providers. Especially in mobile networks, telecommunication operators are natural candidates to support the placement of computation close to the users and allow them to use AI services with small latency. However, if nongeneric models are required, this will result in the migration and placement of user models at the edge. As of now, it is unclear if this is a sound business case. The number of possible applications, e.g., assisted driving, support of the elderly, and on-the-fly translation services, are promising but require a business model to justify such an extension of telecommunication infrastructures. In addition, the multitude of cloud and edge providers, along with their interconnections, makes it challenging to ensure the system reliability that is rightfully expected by consumers. Similarly, ensuring the levels of data protection and trustworthiness that are requested by consumers can be challenging. Providers of edge AI solutions could also use edge AI solutions to improve their service, in which case they may again face challenges with the reliability of their solution. One important aspect is the consideration of consumer applications running on the edge, such that the operations used for improving their operation do not interfere with the regular operation of these applications.

Governmental Perspective

The governmental view of edge AI is multifaceted, encompassing various aspects such as the enforcement of ethical and responsible use, the safeguarding of citizen privacy (echoing societal concerns), the *protection of the intellectual property* of companies, the setup of the necessary infrastructure, the promotion of *interoperability* through common standards, and the monitoring of data exchanges via lawful intercepts. The prioritization of these aspects varies for governments with different focuses. Notably, European nations, already pioneers in privacy regulations like General Data Protection Regulation (GDPR) and security regulations such as the EU Cyber Resilience Act, are likely to emphasize the ethical use of edge AI and the *protection of privacy and security*. We note, finally, that governmental actors can improve the development of edge intelligence through funding and regulations, enabling new services in the respective country.

Developer Perspective

From a developer's viewpoint, the *ease of programming* is crucial for adopting edge AI, particularly for creating distributed applications. Ideally, developers should expend minimal effort in addressing typical edge AI issues such as user and data mobility, distributed coordination, and synchronization. Therefore, to facilitate developer access to edge AI, a programming framework is necessary to simplify the development and configuration of edge AI applications. That includes managing computation and storage resources, automating the watermarking of deployed models, handling the distribution of sensor data, and providing program abstractions for new paradigms like quantum and neuromorphic computing.

Summary and Research Perspective

The research perspective combines all of the aforementioned views into a holistic one, in which future research aims to solve parts of today's problems of edge AI. Several works like Rausch et al.[24] and Nastic et al.[25] have proposed *programming models* for edge AI, such that training and inference can be executed in a decentralized manner; one example is the paradigm of federated learning. While the *locality of data* and the corresponding decentralization can be seen as a positive influence on *data privacy*, trust in edge devices can be limited at times. Thus, there are additional challenges related to the protection of data privacy, for which several ideas are currently being investigated. These include the use of homomorphic encryption, on-device filtering of task-relevant data in hardware-secured execution environments, and research on ensuring trust into edge AI solutions. Research results on these topics can provide valuable inputs to governments regulating edge processing systems.

CURRENT RESEARCH CHALLENGES AND OPPORTUNITIES

Edge AI offers a transformative approach to embedding intelligence into local devices. It is associated with challenges around resource constraints, security and privacy, sustainability, and dealing with the energy crisis. At the same time, it brings significant opportunities in real-time data processing, efficiency, and personalized experiences. The algorithms that make up AI are finding their way into a growing number of excellent services for users. The way this uptake happens and its technical potential was analyzed in several published studies.[2,11,13] This raises a number of issues in understanding the challenges and opportunities of edge AI, from which we highlight the most current and notable ones in the following.

Resource Limitations

Edge devices are characterized by limited computing and storage resources. While cloud-based applications can utilize a variety of computing devices, including CPUs, GPUs, and sometimes field-programmable gate arrays, edge devices commonly contain only a few hardware accelerators that are often tailored for a specific application or use case. In addition, the computing, memory, and storage of edge devices are significantly constrained, limiting the possibility of training and inference even further. This is especially a challenge when it comes to the application of edge AI solutions, as ML models commonly rely on dedicated hardware and require a large volume of memory and storage. In addition, the exchange of data is often critical and limited by the available network bandwidth. Thus, mechanisms need to be developed to limit the amount of exchanged information, not only with central infrastructure, but also between edge devices, e.g., by information-driven prioritization.[27] The training of ML models at the edge is particularly challenging due to these resource limitations, representing an ongoing challenge in the field.

Given that the location of inference is not always predetermined, spanning from powerful centralized devices to resource-constrained edge devices, the necessity for multiple ML models becomes apparent. Each deployment environment comes with its own set of constraints and requirements, whether it is real-time processing on edge devices or comprehensive analysis on robust computational platforms. As a result, developers often need to tailor and optimize models to suit diverse deployment scenarios, ensuring efficiency and effectiveness across the spectrum. Automated mechanisms are required to support this adaptation, such that edge AI solutions can seamlessly integrate into various contexts, catering to the specific needs and constraints of each deployment scenario while maintaining the best possible performance.

Privacy and Trust

Ensuring reliability, security, privacy, and ethical integrity is key to establishing trustworthiness in both edge AI applications and connected systems. This is crucial as edge devices handle sensitive data, and the consequences of breaches can be severe.

Essential to establishing trust is secure processing and storage combined with robust encryption and stringent access controls. AI models must be reliable and accurate, despite the limited resources of edge devices, and robust against adversarial attacks. The use of hardware-supported, trusted execution environments is sometimes considered but comes with its own set of challenges regarding performance and integration. Additionally, transparency and explainability in AI decision-making are increasingly important, especially in critical applications. Compliance with regulations like GDPR, mandating data privacy and security, is also a key aspect of edge AI to be addressed.

Sustainability and Energy Efficiency

The growing need for AI applications emphasizes the importance of creating energy-efficient and sustainable edge AI algorithms. Advanced AI, particularly deep learning, consumes substantial energy,[26] presenting a sustainability challenge. Balancing performance with energy efficiency is crucial for edge AI. While achieving higher levels of accuracy may seem like the ultimate goal, it is imperative to recognize that each incremental improvement in accuracy often demands a substantial increase in energy consumption. This tradeoff becomes particularly apparent in scenarios where ultrahigh accuracy might not be crucial. In such cases, allocating excessive energy resources for marginal gains in accuracy could be inefficient and environmentally unsustainable. Thus, developers and researchers must conscientiously evaluate the necessity of heightened accuracy against the energy footprint it entails.

Another important aspect is the growing importance of renewable energy sources to the energy grid. Since most renewable sources are dependent on environmental conditions (like sunshine for solar cells), there are times, e.g., on a hot summer day with a lot of wind, when power is abundant. While conservative energy usage remains a significant challenge in edge AI, another key challenge is the possibility of performing non-time-critical calculations like model training when there is an energy surplus. Executing these calculations at times of excess power can help balance out spikes in energy generation and compensate for the fluctuating nature of most renewable energy sources. Moreover, distributing the energy demand geographically can help alleviate supply problems faced by large-scale data centers accumulated in certain regions such as Northern Virginia, USA and Amsterdam, The Netherlands. As our energy storage capacities are limited and often inefficient, this can greatly improve the efficiency of the power grid and edge devices.

While energy consumption during operation is an important challenge, so is the production and lifecycle of the deployed edge devices. Designing more durable, upgradeable, and recyclable devices is of critical importance to improve the environmental footprint of edge AI solutions. Additionally, implementing policies to encourage energy-efficient AI and regulating the environmental impact of device manufacturing and disposal is essential.

Programmability and Interoperability

Edge AI involves diverse devices like smartphones, IoT devices, and industrial machinery, each with unique constraints. Creating programmability frameworks for edge AI is challenging due to the need to orchestrate services across this varied hardware efficiently.[17] Developers face the complexity of differing device capabilities in terms of CPU power, GPU availability, memory, and energy consumption.[12] This complexity makes the deployment of services at a large scale such as smart cities a major and already continuing challenge.[14] The lack of standardized tools further complicates development, often requiring the use of incompatible tools and platforms, leading to longer development times and integration issues.

The programmability challenge of edge AI is even further intensified by the need for interoperability, i.e., combining operations on a variety of devices and systems, like sensors, smartphones, and industrial machinery. These devices should work together seamlessly despite different operating systems, software, and hardware. A key issue is the lack of standardized protocols and data formats, making it crucial to develop universal standards for effective communication. Integrating edge AI with existing systems poses challenges because of unsupported software and hardware components. As the number of interconnected devices grows, scalability and easy integration of new devices become important and difficult. Minimizing delays caused by interoperability is crucial in real-time processing scenarios like environmental monitoring, autonomous driving, and industry 4.0. However, managing resources efficiently in this interconnected environment, together with resources available in the continuum, is also a key challenge to be addressed.

To summarize, unified programmability frameworks are essential for deploying edge AI algorithms effectively, ensuring efficient service orchestration, resource management, and device interoperability across the continuum.

Dependability and Resilience

Dependability focuses on the reliability, security, and robustness of AI systems operating on edge computing devices used for AI decision-making. It encompasses ensuring these systems perform consistently and accurately, even in challenging or unpredictable environments.[18] Such systems, crucial in cyber-critical sectors like health care and industrial automation, must always be operational with robust design and effective failover strategies.[16] Developed systems must protect data and AI model integrity against various threats, be capable of handling more data, and accommodate more devices or geographical areas. Systems must autonomously detect and resolve faults and adapt to changing conditions and emerging threats for dependable operation.

In addition to dependability, the resilience of edge AI is pivotal to guarantee its operability at all times. Resilience involves ensuring reliable functioning against offensive security and disruptions under various conditions. Edge devices must be robust against physical challenges like extreme temperatures and mechanical impacts and maintain data integrity and security. Even in poor network conditions, edge systems should either be able to provide reliable connectivity (via alternative communication technologies or robust protocols) or operate offline until connectivity is available again. In addition, these systems need to be fault-tolerant, possibly with backup solutions. AI models should adapt to changing data patterns without needing extensive retraining. As edge AI networks grow, scalability and manageability become key, alongside efficient resource management to handle varying workloads across the continuum.

Measurability

Defining generalized metrics for evaluating performance across the cloud-edge continuum is challenging due to the unique characteristics (e.g., distribution in training and inference, shared resources) and constraints (e.g., resource limitations, real-time requirements) of different edge AI applications and connected systems. This is particularly true for the challenges mentioned earlier, which are currently difficult to measure and quantify. To allow for research in those areas, identifying metrics that accurately measure the development is pivotal.

Additionally, a significant challenge in edge AI involves balancing tradeoffs among accuracy, latency, resource usage, and privacy across this continuum. While simulations or emulations can predict the performance of approaches in certain scenarios, it is essential to verify the validity of developed approaches in real-world testbeds. Developing benchmark evaluation frameworks in such real-world environments and considering real-world use cases remains an open challenge. These benchmarks must rely on generalized metrics to precisely measure and evaluate the unique characteristics of Edge AI.

FUTURE RESEARCH DIRECTIONS

In this section, we identify the most promising research directions: the integration of large language models (LLMs) into edge AI applications, low-latency inference for autonomous vehicles, shifting focus toward energy and privacy in our society, enhancing edge interoperability, and finally advancing trust and security in edge AI systems. We detail each of these directions next.

Integration of LLMs in Edge AI

The integration of LLMs into applications on the edge presents an exciting avenue for future research. LLMs have traditionally been considered too computationally expensive for inference on the edge, relegating them to cloud-based inference. Running them on edge devices introduces a paradigm shift. Increasingly, edge devices are powered with energy-efficient accelerators. For instance, Apple neural engines (ANEs) are available in iPhones and edge-tensor processing units from Google are available as submodules for embedded devices. Running LLMs on these edge accelerators offers the advantage of “free” inference to these companies since the “cost” (primarily: energy consumption) now occurs on end-user devices. This approach could benefit applications with relaxed latency requirements, such as social media platforms, where immediate response is not critical. However, the challenge lies in adapting these computationally intensive models to the constraints of edge devices, including limited processing power and energy efficiency. Classical learning techniques such as distillation, neural architecture search, and systems techniques such as quantization and sparsification are all potential candidates yet “unproven” in their effectiveness. The challenge in evaluating LLMs makes this no easier. Future research should focus not on optimizing LLMs for edge environments but could also lead to innovations in customized hardware that meets the power profiles of the edge.

Edge Computing for Autonomous Agents

Autonomous agents are becoming increasingly important for our society. This includes, e.g., autonomous robots in a smart factory and autonomous vehicles. Even while both the first self-driving vehicles and the first autonomous robots are deployed, these agents currently only function with constant network connectivity, in certain areas, or under certain conditions. But even today, a multitude of sensors, both external and on-board sensors, generate vast amounts of data that could overwhelm traditional internet infrastructures. This data, if shared with low delay, can improve the driving behavior of other agents, allowing for even higher levels of automation. Edge computing allows local processing of data, reducing the need to transfer massive volumes over the network. This not only enhances response times and operational efficiency but also supports real-time decision-making at “internet blind spots” crucial for the reliable operation of the agents. This is not merely a “nice to have,” but essential. The multisensor inputs, perception at different times of the day and in different environments, across diverse weather conditions, and social elements not only advocate for an edge-focused solution but open up new avenues for research in both training and inference aspects on the edge.

Focus on Energy Efficiency and Privacy in Society

The increasing societal awareness of energy consumption and privacy concerns presents a unique opportunity for edge AI. Local processing on edge devices ensures data privacy by keeping sensitive information within the device, thereby guaranteeing data privacy. Moreover, in power-linear systems used on the edge, such as deeply embedded deployments, communication costs (Wi-Fi, Bluetooth, and so on) can be higher than computational expenses. Thus, there is a pressing need for research focused on developing energy-efficient edge AI solutions that balance communication overhead with computational efficiency. This involves exploring energy-aware algorithms, sustainable hardware designs, and optimizing network protocols for energy conservation. Low-power wide-area networks are a promising direction, trading off throughput with power. While this presents one design point in the wide Pareto curve, how to develop solutions that are general-purpose enough to lower production costs, while being tailored enough to support the unique needs of applications is an open research question.

Enhancing Edge AI Interoperability

As edge AI systems become more prevalent, ensuring their scalability and interoperability is a new and unexplored frontier. An open question is how the different AI-enabled edge devices should talk to each other. Beyond the perennial debate of decentralized versus centralized, hub-and-spoke versus circular, one exciting thrust could be on developing standardized protocols and frameworks that enable seamless integration of diverse edge devices and systems. This includes creating universal data formats and communication standards to facilitate efficient interaction and more critically discovery between different types of edge devices, such as sensors, wearables, smartphones, industrial equipment, autonomous vehicles, and so on.

The evolution of edge AI also brings along the need to address interoperability with emerging non-von Neumann architectures (e.g., quantum and neuromorphic computing).[15] It is essential to develop protocols and standards that enable effective communication and collaboration with traditional computing systems. This involves not only the translation of data formats and communication protocols but also the understanding and alignment of the fundamentally different ways in which different non-von Neumann architectures process and interpret data. Neuromorphic computing systems, for instance, mimic the neural structure of the human brain to achieve extreme parallelism and energy efficiency; however, they are event-driven and operate with analog data (spikes). Bridging this gap is crucial in creating a truly interconnected edge AI ecosystem, where devices can leverage the unique strengths of both current and future computing paradigms.

Advancing Trust and Security in Edge AI Systems

Ensuring the trustworthiness and security of edge AI systems are paramount, especially as they become integral to critical infrastructures and personal devices. Future research should focus on developing robust security protocols and encryption methods to protect sensitive data processed at the edge. This includes enhancing the resilience of edge AI systems against cyber threats and ensuring that AI decision-making processes are transparent, explainable, and compliant with regulatory standards like GDPR. This becomes challenging given edge devices sometimes lack trusted environments, which are pivotal for protecting privacy-sensitive data. Addressing these aspects will not only improve the security and reliability of edge AI systems but also foster public trust in their deployment and usage.

CONCLUSION

In this article, we revisited the history and current state of edge AI solutions, ranging from its origin as a combination of edge computing with AI to its current state with decentralized interference and training of AI on resource-constraint edge devices. We highlighted the different challenges and research opportunities of edge AI today, including the perspectives of the relevant stakeholders in the edge AI area. Finally, we envision future research directions for researchers in the field.

REFERENCES

1. Y. Li, H. Wang, and M. Barni, "A survey of deep neural network watermarking techniques," *Neurocomputing*, vol. 461, pp. 171–193, Oct. 2021, doi: 10.1016/j.neucom.2021.07.051.
2. A. Y. Ding et al., "Roadmap for edge AI: A Dagstuhl perspective," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 52, no. 1, pp. 28–33, 2022, doi: 10.1145/3523230.3523235.
3. B. Varghese et al., "Revisiting the arguments for edge computing research," *IEEE Internet Comput.*, vol. 25, no. 5, pp. 36–42, Sep./Oct. 2021, doi: 10.1109/MIC.2021.3093924.
4. W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet*

Things J., vol. 3, no. 5, pp. 637–646, Oct. 2016, doi: 10.1109/JIOT.2016.2579198.

5. L. Lovén et al., "EdgeAI: A vision for distributed, edge-native artificial intelligence in future 6G networks," in *Proc. 6G Wireless Summit*, Levi, Finland, 2019, pp. 1–2.
6. Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proc. IEEE*, vol. 107, no. 8, pp. 1738–1762, Aug. 2019, doi: 10.1109/JPROC.2019.2918951.
7. E. Peltonen et al., "The many faces of edge intelligence," *IEEE Access*, vol. 10, pp. 104,769–104,782, 2022, doi: 10.1109/ACCESS.2022.3210584.
8. O. L. A. López et al., "Energy-sustainable IoT connectivity: Vision, technological enablers, challenges, and future directions," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 2609–2666, 2023, doi: 10.1109/OJCOMS.2023.3323832.
9. J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," 2016, *arXiv:1610.02527*.
10. H. Kokkonen et al., "Autonomy and intelligence in the computing continuum: Challenges, enablers, and future directions for orchestration," 2022, *arXiv:2205.01423*.
11. A. Y. Ding, M. Janssen, and J. Crowcroft, "Trustworthy and sustainable edge AI: A research agenda," in *Proc. 3rd IEEE Int. Conf. Trust, Privacy Secur. Intell. Syst. Appl. (TPS-ISA)*, Atlanta, GA, USA, 2021, pp. 164–172, doi: 10.1109/TPSISA52974.2021.00019.
12. S. G. Patil, P. Jain, P. Dutta, I. Stoica, and J. Gonzalez, "POET: Training neural networks on tiny devices with integrated rematerialization and paging," in *Proc. Int. Conf. Mach. Learn. (ICML)*, PMLR, 2022, pp. 17,573–17,583.
13. R. Singh and S. S. Gill, "Edge AI: A survey," *Internet Things Cyber-Physical Syst.*, vol. 3, pp. 71–92, Jan. 2023, doi: 10.1016/j.iotcps.2023.02.004.
14. B. Cheng, G. Solmaz, F. Cirillo, E. Kovacs, K. Terasawa, and A. Kitazawa, "FogFlow: Easy programming of IoT services over cloud and edges for smart cities," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 696–707, Apr. 2018, doi: 10.1109/JIOT.2017.2747214.
15. D. Kimovski et al., "Beyond von Neumann in the computing continuum: Architectures, applications, and future directions," *IEEE Internet Comput.*, early access, 2023, doi: 10.1109/MIC.2023.3301010.
16. A. Khalil et al., "Dependability: Enablers in 5G campus networks for industry 4.0," in *Proc. 19th Int. Conf. Des. Reliable Commun. Netw. (DRCN)*, Vilanova i la Geltru, Spain, 2023, pp. 1–8, doi: 10.1109/DRCN57075.2023.10108299.
17. X. Guo, A. D. Pimentel, and T. Stefanov, "Automated exploration and implementation of distributed CNN inference at the edge," *IEEE Internet Things J.*, vol. 10, no. 7, pp. 5843–5858, Apr. 2023, doi: 10.1109/JIOT.2023.3237572.
18. X. Guo, A. D. Pimentel, and T. Stefanov, "RobustDiCE: Robust and distributed CNN inference at the edge," in *Proc. 29th Asia South Pacific Des. Automat. Conf. (ASP-DAC)*, Incheon Songdo Convensia, South Korea, Jan. 2024, pp. 26–31, doi: 10.1109/ASP-DAC58780.2024.10473970.
19. E. De Coninck et al., "DIANNE: A modular framework for designing, training and deploying deep neural networks on heterogeneous distributed infrastructure," *J. Syst. Softw.*, vol. 141, no. 7, pp. 52–65, Jul. 2018, doi: 10.1016/j.jss.2018.03.032.
20. S. Leroux, T. Verbelen, P. Simoens, and B. Dhoedt, "Iterative neural networks for adaptive inference on resource-constrained devices," *Neural Comput. Appl.*, vol. 34, no. 13, pp. 10,321–10,336, 2022, doi: 10.1007/s00521-022-06910-5.
21. N. Mohan, L. Corneo, A. Zavodovski, S. Bayhan, W. Wong, and J. Kangasharju, "Pruning edge research with latency shears," in *Proc. 19th ACM Workshop Hot Topics Netw. (HotNets)*, New York, NY, USA: ACM, 2020, pp. 182–189, doi: 10.1145/3422604.3425943.
22. G. Al-Atat, A. Fresa, A. Behera, V. Moothedath, J. Gross, and J. Champati, "The case for hierarchical deep learning inference at the network edge," in *Proc. 1st Int. Workshop Netw. AI Syst. (NetAISys)*, 2023, pp. 1–6, doi: 10.1145/3597062.3597278.
23. V. Moothedath, J. Champati, and J. Gross, "Getting the best out of both worlds: Algorithms for hierarchical inference at the edge," *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 2, pp. 280–297, 2024, doi: 10.1109/TMLCN.2024.3366501.
24. T. Rausch, W. Hummer, V. Muthusamy, A. Rashed, and S. Dustdar, "Towards a serverless platform for edge AI," in *Proc. 2nd USENIX Workshop Hot Topics Edge Comput. (HotEdge)*, 2019, pp. 1–7.
25. S. Nastic, P. Raith, A. Furutanpey, T. Pusztai, and S. Dustdar, "A serverless computing fabric for edge and cloud," in *Proc. IEEE 4th Int. Conf. Cogn. Mach. Intell. (CogMI)*, Piscataway, NJ, USA: IEEE Press, 2022, pp. 1–12, doi: 10.1109/CogMI56440.2022.00011.
26. D. Katare, D. Perino, J. Nurmi, M. Warnier, M. Janssen, and A. Y. Ding, "A survey on approximate edge AI for energy efficient autonomous driving services," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 4, pp. 2714–2754, 4th Quart. 2023, doi: 10.1109/COMST.2023.3302474.
27. D. Bischoff, F. A. Schiegg, D. Schuller, J. Lemke, B. Becker, and T. Meuser, "Prioritizing relevant

information: Decentralized V2X resource allocation for cooperative driving," *IEEE Access*, vol. 9, pp. 135,630–135,656, 2021, doi: 10.1109/ACCESS.2021.3116317.

TOBIAS MEUSER is a sensor researcher and heads the research group of adaptive communication networks in the Chair of Communication Networks at the Technical University of Darmstadt, 64283, Darmstadt, Germany. Contact him at tobias.meuser@kom.tu-darmstadt.de.

LAURI LOVÉN is a postdoctoral researcher and the coordinator of the distributed intelligence strategic research area in the 6G Flagship research program, at the Center for Ubiquitous Computing (UBICOMP), University of Oulu, 90014, Oulu, Finland. Contact him at lauri.loven@oulu.fi.

MONOWAR BHUYAN is an assistant professor of computer science and heads the Cyber Analytics and Learning Group within the Autonomous Distributed Systems Lab at Umeå University, 90187 Umeå, Sweden. Contact him at monowar@cs.umu.se.

SHISHIR G. PATIL is a Ph.D. student in computer science at UC Berkeley, California, Berkeley, CA, 94709, USA. He is affiliated with the Sky Computing Lab (previously RISE), Lab11, and Berkeley AI Research (BAIR). Contact him at shishirpatil@berkeley.edu.

SCHAHRAM DUSTDAR is a full professor of computer science and heads the Research Division of Distributed Systems at Vienna University of Technology, Vienna 1040, Austria. Contact him at dustdar@dsg.tuwien.ac.at.

ATAKAN ARAL is an assistant professor at the Department of Computing Science at Umeå University, 90187, Umeå, Sweden, and a research fellow at the Faculty of Computer Science at the University of Vienna, 1090, Vienna, Austria. Contact him at atakan.aral@umu.se.

SUZAN BAYHAN is an associate professor at Design and Analysis of Communication Systems (DACS) and affiliated with EDGE research center at the University of Twente, 7500 AE, Enschede, The Netherlands. Contact her at s.bayhan@utwente.nl.

CHRISTIAN BECKER is a full professor for computer science and heads the Institute for Parallel and Distributed Systems at the University of Stuttgart, 70569, Stuttgart, Germany. Contact him at christian.becker@ipvs.uni-stuttgart.de.

EYAL DE LARA is a professor of computer science at the University of Toronto, Toronto ON, M5S 1A1, Canada. Contact him at delara@cs.toronto.edu.

AARON YI DING is an associate professor at TU Delft and University of Helsinki (permanent) and leads the Cyber Physical Intelligence (CPI) Lab, 2600 AA, Delft, The Netherlands. Contact him at aaron.ding@tudelft.nl.

JANICK EDINGER is a professor of computer science and head of the research group for Distributed Operating Systems at the University of Hamburg. 22527, Hamburg, Germany. Contact him at janick.edinger@uni-hamburg.de.

JAMES GROSS is a full professor in wireless networking at the School of Electrical Engineering and Computer Science at the Royal Institute of Technology Stockholm (KTH), 100 44, Stockholm, Sweden. Contact him at jamesgr@kth.se.

NITINDER MOHAN is a senior researcher in Chair of Connected Mobility at the Technical University of Munich, 80333, Munich, Germany. Contact him at mohan@in.tum.de.

ANDY D. PIMENTEL is a full professor at University of Amsterdam and chairs the Parallel Computing Systems (PCS) group within the Informatics Institute, 1098 GH, Amsterdam, The Netherlands. Contact him at a.d.pimentel@uva.nl.

ETIENNE RIVIÈRE is a professor of Computer Science and heads the Cloud and Large Scale computing group at UCLouvain, B-1348, Louvain-la-Neuve, Belgium. Contact him at etienne.riviere@uclouvain.be.

HENNING SCHULZRINNE is a professor in the Dept. of Computer Science at Columbia University, New York, NY, 10027, USA. Contact him at schulzrinne@cs.columbia.edu.

PIETER SIMOENS is an assistant professor at the Internet Technology and Data Science Lab at Ghent University-imec, B-9052, Gent, Belgium. Contact him at pieter.simoens@ugent.be.

GÜRKAN SOLMAZ is a senior researcher at NEC Laboratories Europe, 69115, Heidelberg, Germany. Contact him at gurkan.solmaz@neclab.eu.

MICHAEL WELZL is a full professor in the Networks and Distributed Systems group of the Department of Informatics at University of Oslo, 0313, Oslo, Norway. Contact him at michawe@ifi.uio.no.

Get Published in the
IEEE Transactions on Privacy

This fully open access journal is soliciting papers for review.

IEEE Transactions on Privacy serves as a rapid publication forum for groundbreaking articles in the realm of privacy and data protection. Submit a paper and benefit from publishing with the IEEE Computer Society! With over 5 million unique monthly visitors to the IEEE Xplore® and Computer Society digital libraries, your research can benefit from broad distribution to readers in your field.

Submit a Paper Today!

Visit computer.org/tp to learn more.

IEEE COMPUTER SOCIETY

IEEE

EDITORS: Manish Parashar, manish.parashar@utah.edu
David Abramson, david.abramson@uq.edu.au

This article originally appeared in computing in SCIENCE & ENGINEERING vol. 24, no. 2, 2022

DEPARTMENT: TRANSLATIONAL COMPUTER SCIENCE

PyCOMPSs as an Instrument for Translational Computer Science

Rosa M. Badia, Javier Conejero, Jorge Ejarque, Daniele Lezzi, and Francesc Lordan,
Barcelona Supercomputing Center, 08034, Barcelona, Spain

With the advent of distributed computing, the need for frameworks that facilitate its programming and management has also appeared. These tools have typically been used to support the research on application areas that require them. This poses good initial conditions for translational computer science (TCS), although this does not always occur. This article describes our experience with the PyCOMPSs project, a programming model for distributed computing. While it is a research instrument for our team, it has also been applied in multiple real use cases under the umbrella of European Funded projects, or as part of internal projects between various departments at the Barcelona Supercomputing Center. This article illustrates how the authors have engaged in TCS as an underlying research methodology, collecting experiences from three European projects.

Programming parallel and distributed computing systems is a challenging task. Many aspects contribute to it: the complexity of the computing infrastructure with new architectures and heterogeneous devices, the computer and data distribution aspects, or the complexity of the applications that need to leverage the infrastructure computing power.

To address these challenges, multiple groups have been conducting research towards providing programming environments that simplify the development of applications.[1,2]

Among the paradigms to ease the development of parallel applications, a widely supported approach is task-based programming. Based on defining parallelism at the task level, a task may have different granularities: from a few lines of code to a function, to an invocation, to an external binary. Most environments consider identifying data dependencies between tasks and build a directed acyclic graph at execution, with nodes representing tasks and edges data-dependencies between them. The paradigm has proven to be applicable both at node level[3,4] and on distributed computing environments.[5,6]

The PyCOMPSs project[a] was started 15 years ago. One of the project's goals is to produce stable and reliable software that end-user applications can use. This gives us feedback driving new research and developments in the project while at the same time enabling progress in the application research areas.

This article describes our research methodology in programming environments for distributed computing and how translational research in computer science (TCS) has guided our research process.[7]

OVERVIEW

PyCOMPSs[8,9] is a programming environment designed and developed at the Barcelona Supercomputing Center (BSC). Based on the tasks paradigm, the primary goal of this software is to ease the development of parallel applications for distributed computing platforms.

Depending on the granularity of the tasks, the environment can be used to develop traditional task-based applications (fine-grain tasks) or to develop workflows (coarse-grain tasks).

[a] http://compss.bsc.es/, PyCOMPSs is the Python binding of COMPSs. For clarity, in this article, PyCOMPSs is used as a generic term, which includes COMPSs.

1521-9615 © 2022 IEEE
Digital Object Identifier 10.1109/MCSE.2022.3152945
Date of current version 27 May 2022.

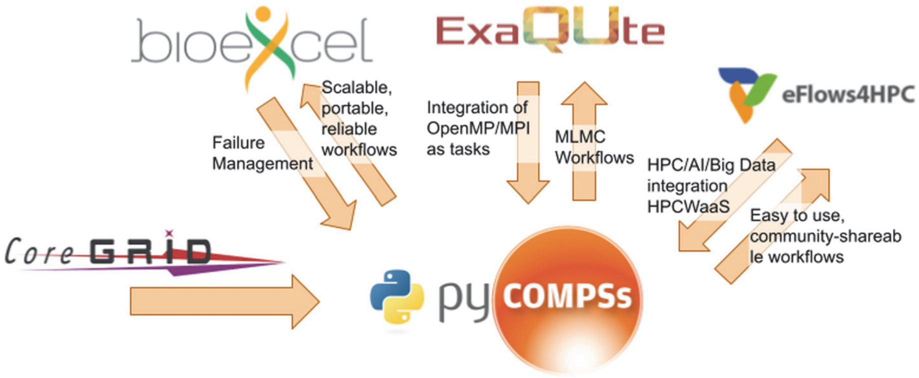

bioexcel
ExaQUte
eFlows4HPC
Scalable, portable, reliable workflows
Failure Management
Integration of OpenMP/MPI as tasks
MLMC Workflows
HPC/AI/Big Data integration HPCWaaS
Easy to use, community-shareab le workflows
CoreGRID
py COMPSs

FIGURE 1. PyCOMPSs and its relation to exemplar EU projects.

With the PyCOMPSs project, we have conducted research on multiple topics, among others: the programming model itself; resource management and its execution in high-performance computing (HPC), clouds, containers; task scheduling; integration with storage and input/output; or convergence of HPC, Big Data, and artificial intelligence.

The project does not target a specific application area, and we have been working in the context of multiple projects with user communities that have provided specific requirements. PyCOMPSs have been leveraged in project use cases in biomedicine, engineering, biodiversity, chemistry, astrophysics, financial, telecommunications, manufacturing, and earth sciences.

Although PyCOMPSs is an academic open-source research project, the code is managed under a continuous integration & continuous deployment process with a testing infrastructure that validates new features before merging. Periodic stable releases are delivered twice per year. We perform multiple training activities, and the team members provide support under a best-effort approach.

BSC is proud of having projects that live beyond their funding schemes. For example, BSC's performance tools have been developed for more than 25 years and PyCOMPSs for around 15 years now. While specific developments from basic research projects may be only done as prototype versions, most of those from funded projects are integrated into the official versions.

TRANSLATION PROCESS

TCS in European (EU) Funded Projects

From its early times when COMPSs was started in the CoreGRID[b] project, the framework has been involved in more than 25 EU-funded projects. While Abramson and Parashar[7] identified general shortcomings in existing funding schemes for supporting TCS, some EU funding schemes support projects that include translational computing in some sense.

The European Commission (EC) not only funds basic research programs, such as the prestigious ERC or the Marie Curie award. For example, the current H2020 program considers the following two types of schemes: research and innovation actions (RIAs) and innovation actions (IAs).

RIAs fund more research-oriented activities but still expect industry involvement to explore the possible industrial feasibility of the research results. In IAs, funding focuses on closer-to-the-market activities, including prototyping, testing, demonstrating, piloting, etc. In addition, the "HPC centres of excellence" (CoE) address scientific applications and user communities running application codes or large-scale workloads seeking an extreme scaling performance.

These are three examples of the funding schemes in Europe. All of them are collaborative in nature, expecting a consortium that consists of universities, research institutes, small and medium enterprises (SMEs), and large companies. The partners can play different roles in the consortium: research or technology provider, application provider, end-user, etc. In many cases, industry plays the end-user role, providing use cases that are somehow prototyped and evaluated in the project. This does not mandate TCS, but from the authors' point of view, it may help in the process.

The EC indicates in the call text the expected impact of the funded projects, for example, indicating the level of innovation and productivity enhancement that is sought.

In this article, we will focus on three exemplar EU funding experiences on which we have been involved

[b] http://coregrid.ercim.eu/

with PyCOMPSs: The BioExcel CoE, the ExaQUte RIA project, and the eFlows4HPC IA project (see Figure 1).

BioExcel CoE

BioExcel[c] is the HPC European Centre of Excellence for Computational Biomolecular Research. BioExcel's mission is to provide applications, tools, support, and networking opportunities to address grand scientific challenges that fully exploit the power of large e-infrastructures.

Our team focuses on a collaboration with the Institute for Research in BioMedicine toward the design and development of the BioExcel Building Blocks (BioBBs) library.[10]

> *WHILE THE BIOBB DEVELOPMENT TEAM HAS CONDUCTED ITS RESEARCH ON WORKFLOWS FOR BIOMOLECULAR RESEARCH, THE PYCOMPSS TEAM HAS PERFORMED MORE GENERIC RESEARCH ON THE RESOURCE MANAGEMENT AND EXECUTION OF LARGE WORKFLOWS IN HPC SYSTEMS.*

Each BioBB is a Python wrapper on top of biomolecular simulation tools. The building blocks share a unique syntax, requiring input files, output files, and input parameters, irrespective of the wrapped program. Workflows assembled by composing multiple building blocks and packaged in a single Python script with a defined Conda environment to be shared and reproduced. Multiple workflow engines can enact BioBB workflows, specifically PyCOMPSs, when targeting HPC environments.

During the development of the BioBB library, the BioBB development team and the PyCOMPSs team have been working in collaboration. While the BioBB development team has conducted its research on workflows for biomolecular research, the PyCOMPSs team has performed more generic research on the resource management and execution of large workflows in HPC systems.

Multiple specific research topics in the PyCOMPSs team have arisen thanks to the BioBB workflow's requirements. Some of these topics were related to fault tolerance: management of application failures at task-level, support to application restart, and application checkpointing. For example, the original behavior of PyCOMPSs applications was to safely terminate the whole application if a task error was detected. The BioBB developers indicated to the PyCOMPSs team that such behavior was too conservative and would like to enable the workflows to continue the execution even when some tasks failed. We extended the PyCOMPSs syntax to enable the developer to indicate the desired behavior if an error occurred in a task.[11] For example, this interface allows you to tell the runtime to ignore individual tasks' errors and continue execution and cancel the execution of erroneous task successors.

The BioBB library is offered to the user community through tutorials, as ready-to-use workflows, or as source code to build new workflows. An example of the level of readiness of the BioBB workflows has been demonstrated with the SARS-CoV-2 emergency. In early 2020, the two teams worked together to define a set of pre-exascale workflows. However, the pandemic changed the priorities toward research questions related to the virus's evolutionary path or the different human sensitivity reactions. The BioBB and their workflows were quickly adapted to answer these questions, and they were run in the MareNostrum 4 supercomputer at BSC.

ExaQUte Project

The ExaQUte project aimed at constructing a framework to enable uncertainty quantification and optimization in complex engineering problems, using computational simulations on exascale systems. The project ran from May 2018 to the end of 2021.

The project was based on multilevel Monte Carlo (MLMC) to enable a large number of stochastic variables. The MLMC algorithm is implemented with the xMC library,[d] which explores multiple simulations of the Kratos multiphysics software.[e] Kratos is fed with meshes of different characteristics defined with the ParMmg software.[f] The whole framework is integrated with Python scripts annotated with PyCOMPSs decorators to support parallelism and distributed computing.

Concerning PyCOMPSs, the main goal was the extension of existing task schedulers to extract the parallelism of the MLMC algorithm and to support distinct levels of complexity of the tasks (OpenMP and MPI tasks), which implies different duration and

[c] https://www.bioexcel.eu

[d] https://doi.org/10.5281/zenodo.3235832

[e] https://www.cimne.com/kratos/

[f] https://www.mmgtools.org/

different amounts of resources used by each task. A critical paradigm to support was the relaxation of global synchronizations required between simulations belonging to different levels of the MLMC algorithm.[12] In addition, the support for MPI tasks was very naive at that time and required extensions to support more complex data layouts.

The research in the project was conducted at multiple levels, with subobjectives, mapping to the components mentioned before (xMC, Kratos, ParMmg, and PyCOMPSs). While the components were providing requirements to others, the main driver was a final user application aimed at robust optimization of structures subject to wind action.

Although ExaQUte involved a small number of partners, it had the typical consortium structure, with an end-user from industry, research and technology providers, and two supercomputer centers providing infrastructure. The final user was the SME str.ucture, whose main engineering activities lie around the industrial application of advanced computer-aided simulation methods on parallel HPC platforms. The company has leveraged the research performed in the project to study different use cases of its interest, like the assessment of wind-induced galloping instability of cable cars or the wind effects on large span bridges. This second research was done in collaboration with German structural engineering companies.

eFlows4HPC Project

With the experience obtained in previous projects, our research group proposed the eFlows4HPC project. The technical objective of the project goes beyond PyCOMPSs and proposes a whole software stack for the development of workflows that involve HPC simulation and modeling, artificial intelligence, and Big Data analytics.

The project aims to demonstrate through three application Pillars of high industrial and social relevance how the realization of forthcoming efficient HPC and data-centric applications can be developed adopting new workflow technologies. It will integrate existing workflow interfaces, programming models, artificial intelligence, and data analytics libraries to provide a uniform, easy-to-use platform that enables the exploitation of future large-scale systems. eFlows4HPC also contributes with the HPC Workflows as a Service idea to widen newcomers' access to HPC, and in general, to simplify the deployment and execution of complex workflows in HPC systems, providing mechanisms to enable the sharing, reuse, and reproducibility of complex workflows.

The eFlows4HPC project involves three application pillars to define complex workflows based on the project technologies. While validating the technologies, these workflows are also a vehicle for research on the pillar topics: digital twins for manufacturing, climate prediction, and urgent computing for natural hazards.

Another goal is the user communities adopting the project solutions to enable impact on industrial cases and their exploitation in future HPC systems. To reinforce the feedback to communities, the Centers of Excellence ChEESE,[g] ESiWACE,[h] and EXCELLERAT[i] are involved in the project, in coordination with the Focus CoE.[j] To this end, the project activities include organizing workshops and training schools that will transfer the project methodologies and results to the relevant CoEs and industrial communities, contributing to the reduction of skills gaps in Europe related to HPC and workflows development.

IMPACT

This article has described several cases where PyCOMPSs is used in translational computer science research. For the activities in the BioExcel project, the impact of the translational process is evident for us. All the stakeholders involved in the activities have benefited from it. PyCOMPSs has been enhanced with new features, which were fresh and exciting enough to imply innovative research contributions.[11] A significant plus for our team is that these new features were helpful for the BioBB workflows and have fostered the research and development of workflows for molecular dynamics. Finally, these activities impact the user community, with new workflows available for their research, and have been helpful for emerging research activities, such as the COVID-19 investigations.

Similar conclusions can be derived for the ExaQUte project. The project put together a whole software stack integrating components from different partners. Beyond the research on programming models for distributed computing, other partners conducted research on MLMC or simulation of structures, to name a few. Multiple requirement-feedback loops were exchanged between the consortium partners. A small spin-off project, EdgeTwinsHPC,[k] was funded to explore the viability of HPC software to generate digital twins that run on the edge. Part of this research is also conducted in Pillar I of the eFlows4HPC project.

We hope that in a couple of years we will also be able to describe similar success stories for the eFlows4HPC

[g] https://www.cheese-coe.eu
[h] https://www.esiwace.eu
[i] https://www.excellerat.eu
[j] https://www.hpccoe.eu/about/
[k] https://www.edgetwins.eu

project which, from our point of view, has been driven by translational computing methodology.

LESSONS LEARNED

Our experience shows that research translation is an iterative process in which new ideas appear after some work. The implementation of the new ideas enables further developments that can provide fresh ideas, and so on. This turns out to have benefits to the different teams involved in the translation process.

One of the challenges of translational research is the multidisciplinary aspect that requires some adjustments in the way of working. The progress sometimes seems to be slow due to different communication languages, priorities, or skills (i.e., what is easy for one group looks very difficult for another). One lesson learned in this sense is that patience is essential. One should not get disappointed if the initial results do not look as expected.

Another lesson that we have learned is that at BSC, we are privileged because we have multiple research departments working on different topics in the same place. While working with cross-disciplinary teams requires some effort, the impact and results exceed the investment. In this sense, our recommendation would be to foster the foundation of interdisciplinary centers and encourage their groups to work together in joint projects.

We recognize a trade-off between the cost of delivering stable software, usable for others, and the number of research publications produced. The effort is not always recognized, at least in the short term. Nevertheless, it is worth it: besides offering exciting tools to the community, starting a new research topic from stable, reliable software is a powerful beginning. The group has been able to deal with this concern and at the same time enjoy enough flexibility and creativity to conduct research on new topics.

CONCLUSION

This article has presented the authors' view of the translational computer science methodology involved with the PyCOMPSs project. PyCOMPSs is a parallel programming model for distributed computing developed at BSC. Since its infancy, it has been partially supported by multiple EU-funded projects. Instead of proposing a brand new independent idea for each project, which is developed as a prototype discontinued after the funding period, the approach has been to allow the project to continue alive thanks to multiple cycles of funding.

In addition, we have leveraged the privileged situation at BSC with multidisciplinary research departments and the collaborative nature of most EU funding schemes to implement a TCS approach.

ACKNOWLEDGMENTS

This work was supported in part by Spanish Government under Contract TIN2015-65316-P, in part by the Generalitat de Catalunya under Contract 2014-SGR-1051, and in part by the European Commission's Horizon 2020 Framework program through BioExcel Center of Excellence under Contract 823830 and Contract 675728, in part by the ExaQUte Project under Contract 800898, in part by the European High-Performance Computing Joint Undertaking (JU) under Grant 955558, in part by the MCIN/AEI/10.13039/501100011033, and in part by the European Union NextGenerationEU/PRTR.

REFERENCES

1. E. Deelman *et al.*, "Pegasus: A framework for mapping complex scientific workflows onto distributed systems," *Sci. Program.*, vol. 13, no. 3, pp. 219–237, 2005, doi: 10.1155/2005/128026.
2. Y. Babuji *et al.*, "Parsl: Pervasive parallel programming in python," in *Proc. 28th Int. Symp. High- Perform. Parallel Distrib. Comput.*, 2019, pp. 25–36, doi: 10.1145/3307681.3325400.
3. A. Duran *et al.*, "OmpSs: A proposal for programming heterogeneous multi-core architectures," *Parallel Process. Lett.*, vol. 21, no. 2, pp. 173–193, 2011, doi: 10.1142/S0129626411000151.
4. P. Bellens, J. M. Perez, R. M. Badia, and J. Labarta, "Cellss: A programming model for the cell be architecture," in *Proc. ACM/IEEE Conf. Supercomput.*, 2006, pp. 86–96, doi: 10.1145/1188455.1188546.
5. R. M. Badia, J. Labarta, R. Sirvent, J. M. Pérez, J. M. Cela, and R. Grima, "Programming grid applications with grid superscalar," *J. Grid Comput.*, vol. 1, no. 2, pp. 151–170, 2003, doi: 10.1023/B:GRID.0000024072.93701.f3.
6. E. Tejedor and R. M. Badia, "Comp superscalar: Bringing grid superscalar and GCM together," in *Proc. 8th IEEE Int. Symp. Cluster Comput. Grid*, 2008, pp. 185–193, doi: 10.1109/CCGRID.2008.104.
7. D. Abramson and M. Parashar, "Translational research in computer science," *Computer*, vol. 52, no. 9, pp. 16–23, 2019, doi: 10.1109/MC.2019.2925650.
8. F. Lordan *et al.*, "ServiceSs: An interoperable programming framework for the cloud," *J. Grid Comput.*, vol. 12, no. 1, pp. 67–91, 2014, doi: 10.1007/s10723-013-9272-5.
9. E. Tejedor *et al.*, "PyCOMPSs: Parallel computational workflows in Python," *Int. J. High Perform. Comput. Appl.*, vol. 31, pp. 66–82, 2017, doi: 10.1177/1094342015594678.

10. P. Andrio *et al.*, “Bioexcel building blocks, a software library for interoperable biomolecular simulation workflows,” *Sci. Data*, vol. 6, no. 1, pp. 1–8, 2019, doi: 10.1038/s41597-019-0177-4.
11. J. Ejarque, M. Bertran, J. Á. Cid-Fuentes, J. Conejero, and R. M. Badia, “Managing failures in task-based parallel workflows in distributed computing environments,” in *Proc. Eur. Conf. Parallel Process.*, 2020, pp. 411–425, doi: 10.1007/978-3-030-57675-2_26.
12. R. Tosi, R. Amela, R. M. Badia, and R. Rossi, “A parallel dynamic asynchronous framework for uncertainty quantification by hierarchical Monte Carlo algorithms,” *J. Sci. Comput.*, vol. 89, no. 1, pp. 1–25, 2021, doi: 10.1007/s10915-021-01598-6.

ROSA M. BADIA is a group manager with the Computer Science Department, Barcelona Supercomputing Center, Barcelona, Spain. Her research interests include parallel programming models, distributed computing, workflow environments, and convergence of AI and HPC. Badia received the Ph.D. degree in computer science from the Universitat Politecnica de Catalunya, Barcelona, Spain. She is an ACM distinguished member and a Member of IEEE. Contact her at rosa.m.badia@bsc.es.

JAVIER CONEJERO is a senior researcher with the Computer Science Department, Barcelona Supercomputing Center (BSC), Barcelona, Spain. His research interests include QoS, development paradigms, parallel and distributed computation, HPC, and cloud computing. Conejero received the Ph.D. degree in advanced computer technologies from the University of Castilla-La Mancha (UCLM), Ciudad Real, Spain. Contact him at javier.conejero@bsc.es.

JORGE EJARQUE is a senior researcher with the Workflows and Distributed Computing Group, Barcelona Supercomputing Center, Barcelona, Spain. His main research interests include the development of parallel programming models for distributed computing platforms contributing to several international R&D projects. Ejarque received the Ph.D. degree in computer science form the Technical University of Catalonia, Barcelona. Contact him at jorge.ejarque@bsc.es.

DANIELE LEZZI is a senior researcher with the Computer Sciences Department, Barcelona Supercomputing Center, Barcelona, Spain. His research interests include high performance, distributed, grid and cloud computing and programming models. Lezzi received the Ph.D. degree in computer engineering from the University of Salento, Lecce, Italy. He is a member of ACM. Contact him at daniele.lezzi@bsc.es.

FRANCESC LORDAN is a researcher with the Computer Science Department, Barcelona Supercomputing Center (BSC-CNS), Barcelona, Spain. His research interests include programming models, distributed computing, and IoT-Edge-Cloud Continuum. Lordan received the Ph.D. degree in computer architecture from the Universitat Politecnic de Catalunya, Barcelona. Contact him at francesc.lordan@bsc.es.

EDITORS: Timothy Jones, University of Cambridge, U.K., timothy.jones@cl.cam.ac.uk
Robert Mullins, University of Cambridge, U.K., robert.mullins@cl.cam.ac.uk

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

DEPARTMENT: COMPUTING ARCHITECTURES

Distributed Quantum Computing via Integrating Quantum and Classical Computing

Wei Tang and Margaret Martonosi, *Princeton University*

As quantum computing confronts scalability challenges, distributed hybrid QPU–CPU techniques emerge as a crucial solution. These techniques distribute quantum algorithms across quantum and classical computing resources to surpass the computational reach of either one alone.

As we navigate an era where the insatiable demand for computing power is increasingly outpacing the capabilities of traditional computing, the slowing progress of Moore's law poses a formidable challenge. At this critical juncture, quantum computing (QC) emerges with the potential to greatly extend computing capabilities in key application domains.

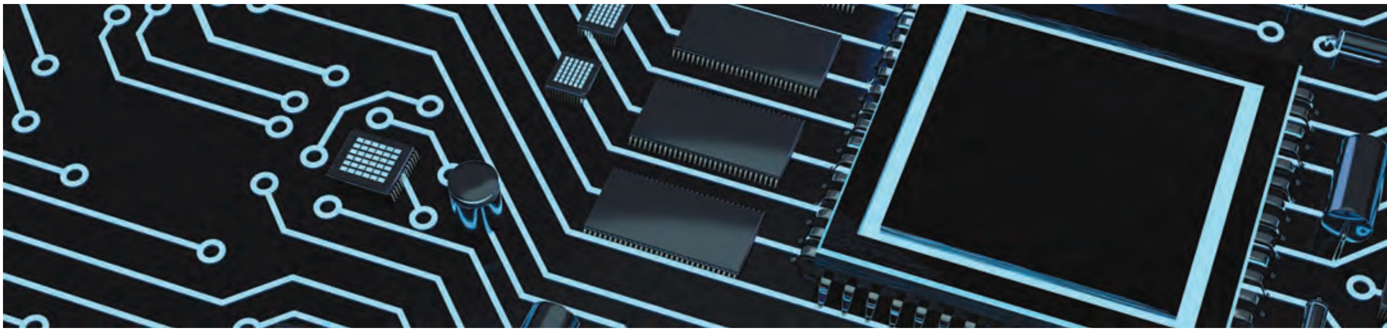
Unlike classical computing operating on binary information, QC rests on quantum bits, or *qubits*, exploiting the nonintuitive yet powerful properties of quantum mechanics, such as entanglement and superposition. This gives QC the potential to perform complex calculations at speeds unattainable by its classical counterparts, toward solving problems considered intractable today. However, scaling up QC systems to these levels involves overcoming substantial engineering and scientific challenges. Current efforts in QC primarily concentrate on enhancing the performance of a single quantum processing unit (QPU). The goal of these efforts is to increase both QPU size and precision, paving the way for QC to tackle real-world applications that were once thought beyond reach.

By contrast, our work seeks to exploit distributed computing that hybridizes quantum and classical approaches. In today's commercial world, distributed and parallel classical computing is not just conceptual; it's an integral part of our daily tech interactions. Distributed classical computing divides complex tasks across multiple computers for simultaneous processing. Unlike the case of a single supercomputer handling all of the tasks, distributed computing spreads these tasks across several, or sometimes thousands of, CPUs, GPUs, and other units. Video games, high-resolution video editing, and artificial intelligence are just a few examples that leverage the collective power of numerous CPUs and GPUs to tackle tasks that would be impossible for a single machine.

The aim of this article is to provide an in-depth exploration of the emerging field of *distributed QC*, particularly through the lens of integrating quantum and classical computing paradigms. In classical distributed computing, data parallelism and model parallelism are two key strategies for processing large or complex tasks. Data parallelism divides a large dataset into smaller chunks, distributing them across multiple nodes for parallel processing. Each node works on its data segment independently, necessitating the initial distribution of these data parts and possibly aggregating the results later. Another approach, model parallelism, involves splitting a complex model, such as a neural network, across nodes, with each working on a different part. This strategy requires a continuous exchange of intermediate results among nodes for collaborative processing.

Digital Object Identifier 10.1109/MC.2024.3360569
Date of current version: 5 April 2024

2469-7087/25 © 2025 IEEE



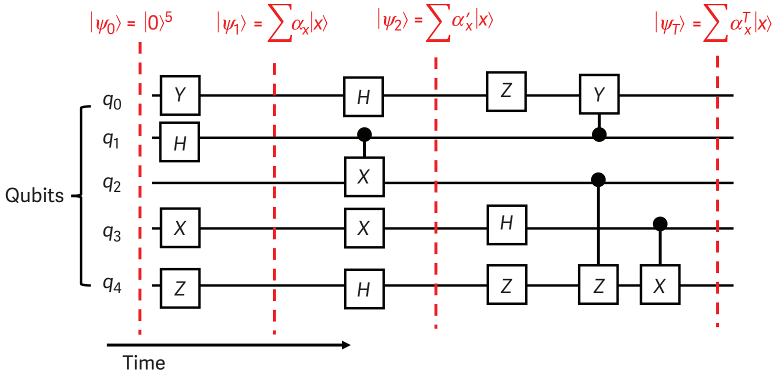

$|\psi_0\rangle = |0\rangle^5$
$|\psi_1\rangle = \sum \alpha_x |x\rangle$
$|\psi_2\rangle = \sum \alpha'_x |x\rangle$
$|\psi_T\rangle = \sum \alpha^T_x |x\rangle$
Qubits
q_0
q_1
q_2
q_3
q_4
Y
H
X
X
Z
H
X
X
H
Z
H
Z
Y
Z
X
Time

FIGURE 1. Example of a five-qubit quantum circuit. Final measurements sample and compute the α^T_x coefficients, or amplitudes, for each possible quantum state, representing a target solution.

At first glance, distributing QC tasks across multiple nodes seems a straightforward extension of classical distributed computing. However, the field faces a unique challenge rooted in the fundamental laws of quantum physics. The quantum no-cloning theorem,[5] a cornerstone principle in quantum mechanics, dictates that it is impossible to create an exact copy of an arbitrary unknown quantum state. This prohibition against duplicating quantum data presents a significant obstacle in developing distributed QC systems as it contradicts the typical data-sharing methodologies employed in classical distributed computing. This article plots a way forward for navigating this challenge, exploring innovative approaches to distribute quantum tasks without copying quantum data. We examine the intersection of quantum and classical computing techniques, shedding light on how this hybrid model can potentially unlock new capabilities and applications in the QC landscape.

BACKGROUND

Unlike classical computers, which use binary 0 or 1 bits as the smallest unit of data, quantum computers use quantum bits, or "qubits." Qubits have the unique ability to exist in multiple states at once, thanks to a quantum phenomenon known as superposition. Imagine a coin that can spin in the air without ever landing—its state is a probabilistic superposition of the heads and tails outcomes it might land as. Another principle is entanglement, where two qubits become linked in such a way that the state of one can instantly affect the state of another, regardless of the distance between them. Knowing whether one coin lands as heads or tails offers the observer information about the state of another entangled coin, even if distant. These properties are not observable at the scale of objects like coins, but they are real physical properties that are measurable at the atomic scale. By exploiting these properties, quantum computers can perform complex calculations at incredible scale, potentially solving problems that are currently intractable for classical computers.

At its core, a quantum program is expressed as a circuit composed of a sequence of quantum operations, known as *gates*, which act on the qubits. These gates, which can be single-qubit or multiqubit operations, play a crucial role in manipulating the states of the qubits. Figure 1 shows an example quantum circuit with five qubits. Each horizontal line denotes a qubit. Boxes incident on a single qubit wire are single-qubit quantum gates, which operate on that qubit. Boxes incident on two qubit wires are two-qubit quantum gates, which operate on both of them.

As the circuit progresses, gates are applied in sequence, altering the collective quantum states of the qubits. This process starts from an initial quantum state on the left-hand side and evolves toward the desired final quantum state on the right. To control this process, sophisticated control electronics are used. These electronics manage the timing, order, and type of quantum gates applied to the qubits. They often include precision timing equipment and can include microwave pulse generators for systems like superconducting qubits, or laser systems for ion-trap qubits.

The process terminates with measuring the qubits, which produces a classical binary string—a phenomenon known as the collapse of the quantum state. Because quantum states and operations are probabilistic in nature, the quantum circuit is executed multiple times. This repetition allows for the accumulation of statistical data, reflecting the final amplitude coefficients, denoted as α_x^T, for all possible quantum states. It is through this meticulous and repeated application of quantum gates and measurements that the quantum circuit unveils the solution encoded in the target quantum state. The illustrated circuit requires a QPU with at least five "good enough qubits" and "accurate-enough operations" to execute all of the quantum gates before too many errors accumulate to produce useful results.

Toward distributed QC

Traditionally, QC has primarily concentrated on the development and optimization of a single QPU, with the aspiration that it will eventually become sufficiently large and accurate to execute complex quantum circuits of practical significance. However, this approach encounters a formidable scalability challenge: many practical quantum applications require thousands of high-quality qubits. These are either implemented logically as error-corrected versions of millions of noisy and error-prone physical qubits, *or* the underlying physical qubits must have sufficient fidelity to avoid the need for error correction. Either way, the scale and complexity introduces significant engineering obstacles, making the realization of practical QC applications a daunting task. Against this backdrop, the role of distributed QC—which harnesses the collective power of multiple QPUs to share and process quantum workloads—has become increasingly crucial. By adopting a distributed framework, the scaling challenges of a single-QPU system can be substantially mitigated, greatly enhancing the potential scope and impact of QC.

TOWARD DISTRIBUTED QC: WHAT IS CIRCUIT CUTTING?

Achieving the goals of QC requires a practical solution to its challenges by breaking down large, complex quantum circuits into smaller, more manageable subcircuits. Circuit cutting is an innovative technique[2] that offers practical approaches for the individual subcircuits to be processed on different QPUs in parallel, and for the classical reconstruction phase that has traditionally stymied QC circuit-cutting techniques.

At the heart of circuit cutting is the concept of decomposition: it essentially involves identifying specific points, known as cut points, within a quantum circuit and then decomposing the complex quantum states at these points into a series of classical components based on a mathematical framework known as *Pauli bases*. Once each subcircuit is run on a QPU, the original, full quantum state must be reconstructed through classical postprocessing, where the results of the separate subcircuits are combined in a specific and compute-intensive way. Naive implementations of classical reconstruction involve matrix multiplications and scale exponentially with factors like qubit state and cut points. Our work improves on these naive reconstruction approaches. By enabling QC circuit decomposition and by helping the subsequent reassembly of quantum tasks to be more tractable, circuit cutting effectively bridges the gap between the current capabilities of quantum hardware and the demands of complex quantum computations, making it a key technique in advancing the field of QC.

A QC circuit-cutting example

Figure 2 illustrates the process of circuit cutting using the straightforward quantum circuit example from Figure 1. In this example, circuit cutting is applied by making a strategic cut, denoted by a red cross, effectively dividing the original circuit into two smaller subcircuits.

The real power of circuit cutting is showcased in the next step, where these subcircuits are assigned to multiple three-qubit QPUs. These three-qubit QPUs only need to support the smaller subcircuits, hence placing fewer requirements on hardware quality. In addition, this approach introduces flexibility and efficiency as these QPUs can operate the subcircuits

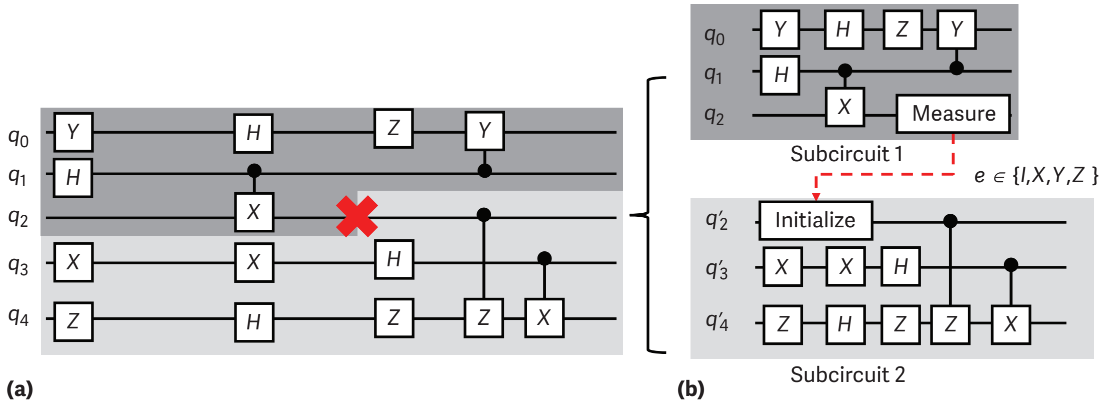


FIGURE 2. Example of cutting a five-qubit quantum circuit with one cut that divides it into two smaller subcircuits. (a) The red cross indicates the cutting point. Subcircuit 1 is shaded dark, and subcircuit 2 is shaded light. (b) Subcircuit 1 with measurements and subcircuit 2 with initialization in each one of the Pauli bases. The two subcircuits require no quantum communications and are executed independently in any order on multiple three-qubit QPUs.

independently and in parallel, without the need for direct quantum communication between them. This independence is possible because the subcircuits are entirely decoupled by the cut.

What does a cut actually mean? At the red X, circuit cutting requires us to mathematically decompose the quantum state at the cut point into its four Pauli bases in {*I, X, Y, Z*}. This mathematical decomposition then allows classical computing to reconstruct the quantum state after QPU execution. Circuit cutting is thus characterized by making vertical cuts across the qubit wires, effectively segmenting a large quantum circuit into several smaller parts. In more complex scenarios, a large circuit might be divided using multiple cuts, further breaking down the computational task into even smaller subcircuits. This technique not only makes quantum computations more feasible on current quantum hardware but also significantly expands the range of problems that can be tackled using available QC resources.

Classical reconstruction postprocessing

As previously noted, straightforward or naive classical reconstruction methods are computationally expensive, bordering on intractable. For example, early proposals for circuit cutting,[2,4] while straightforward in approach and feasible to implement, involve a series of computationally intensive steps. Consider the method of Tang et al.,[4] which at its core requires the computation of the tensor product of the outputs from each subcircuit corresponding to a specific Pauli basis. This process is not a one-off calculation but must be repeated *for every possible combination of Pauli bases*. Once these tensor products for each Pauli basis combination are calculated, the next step involves summing up all of these intermediate results to achieve the final output. The complexity of this method becomes evident when considering the number of Pauli bases involved: with four Pauli bases associated with each cut, the total number of tensor product calculations needed grows exponentially with the addition of each cut.

Challenges

Circuit cutting hence encounters two significant challenges. The first challenge lies in the scalability of the approach. The initial theoretical proposal[2] for circuit cutting included a reconstruction formula for reassembling the final quantum state from the subcircuits that scales exponentially with the number of cuts made in the circuit. This exponential scaling poses a significant obstacle to the practical implementation of circuit cutting, especially for very large and complex quantum circuits.

Second, identifying optimal cut points within large quantum circuits is a complex task. The process involves not just splitting the circuit, but doing so in a manner that ensures each resulting subcircuit is computationally manageable and capable of being

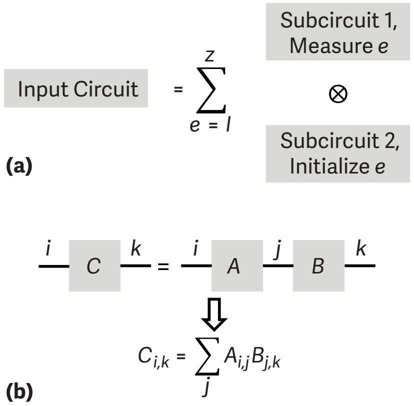

Input Circuit
=
$\sum_{e=l}^{z}$
Subcircuit 1, Measure *e*
⊗
Subcircuit 2, Initialize *e*
(a)
i C *k* = *i* A *j* B *k*
$C_{i,k} = \sum_j A_{i,j}B_{j,k}$
(b)

FIGURE 3. Reconstructing two subcircuits is equivalent to a pairwise tensor contraction. (a) Reconstructing a pair of subcircuits means multiplying and summing over the cut edges in between. (b) A pairwise tensor contraction with one shared index *j*, which is the inner dimension being contracted. *i, k* are the outer dimensions of the resulting big tensor *C*.

processed independently. This requires a careful balance between the complexity of individual subcircuits and the overall efficiency of computation.

The current state of circuit cutting

The first comprehensive implementation of circuit cutting[4] marks a significant advancement in this field by introducing an automated solver algorithm. This algorithm is designed to determine the minimal number of cuts necessary to divide a large quantum circuit into smaller subcircuits, which can then be processed on available smaller scale QPUs. To achieve this, the problem of finding these optimal cuts is formulated as a mixed-integer programming problem, enabling a more systematic and efficient approach to circuit segmentation.

A key aspect of this implementation is its adoption of a relatively straightforward method for the classical reconstruction of the quantum state postcomputation. However, minimizing the number of cuts becomes a critical objective. The reason is that beyond a certain circuit complexity level in terms of size or connectivity, the time and resources required for classical postprocessing overshadow the benefits gained from dividing the circuit, turning it into a computational bottleneck. The success of this implementation, therefore, hinges on striking a delicate balance—optimizing the circuit to fit smaller QPUs through the fewest possible cuts while keeping the postprocessing demands manageable to make circuit cutting a viable and practical approach in QC. Specifically, the work of Tang et al.[4] demonstrates running circuits up to 100 qubits.

Circuit cutting with tensor contraction

The state-of-the-art circuit-cutting technique[3] proposes to integrate distributed QC with classical tensor networks to exponentially improve the postprocessing process, hence eliminating the major obstacle for practical circuit cutting. Tensor networks have been widely used in classical simulations of quantum systems.[1] At its core, a tensor network consists of tensors (multidimensional arrays of numbers) connected by edges, where each edge represents a shared dimension or index between the tensors.

Tensor network contraction is a computational process in which the tensors in a network are systematically combined, or "contracted," according to specific rules. This contraction involves summing over shared indices or dimensions between connected tensors, effectively reducing the network into a single tensor to represent the results of the original network. On the other hand, classical reconstruction for circuit cutting also involves multiplying the subcircuit results across their shared cut qubit wires and taking the summation. Figure 3 shows the mathematical equivalence between the two processes.

Utilizing tensor network contraction in circuit cutting offers an exponential computational advantage over the simple brute-force reconstruction method, primarily because of its efficiency in managing high-dimensional data. In brute-force reconstruction, the process involves calculating and summing the tensor products for each possible combination of Pauli bases across all cuts, leading to an exponential increase in computations with the addition of each cut. By contrast, tensor networks contract the subcircuit tensors along their shared dimensions; this method effectively consolidates the network, step by step, into a single tensor.

This approach dramatically reduces the number of operations required as it eliminates the need to compute every possible combination of tensor products independently. Consequently, tensor network contraction transforms what would be an exponential problem in the brute-force method into a much more tractable one, providing a scalable and efficient way to reconstruct the quantum state in circuit-cutting scenarios, particularly those involving a large number of cuts.

Figure 4 provides an overview of the practical application of circuit cutting, showcasing its runtime across a diverse range of QC benchmarks. These benchmarks include tasks from quantum optimization algorithms and entangled states generation to key subroutines in quantum number factoring algorithms, all of which are pivotal in demonstrating the real-world applicability of QC. In these experiments, each benchmark circuit is constrained to a maximum of half the qubits and gates compared to its original, uncut counterpart, with the largest circuits tested on a 100-qubit QPU for benchmarks designed for up to 200 qubits. Notably, using tensor networks in circuit cutting (ScaleQC) is more than 1 billion times less classical post-processing overhead than brute force (CutQC).

This approach highlights the significant role of circuit cutting in enabling quantum computations that were previously unattainable because of hardware limitations. Without the use of circuit cutting, existing QC methods are restricted to executing quantum programs of considerably smaller scale. Moreover, the complexity and size of these benchmarks far exceed the capabilities of classical simulators; underlining the crucial enhancement that circuit cutting brings to the field of QC, particularly in bridging the gap between current quantum hardware limitations and the demands of advanced quantum algorithms.

Industry acceptance

The industry's embrace of circuit-cutting technology is underscored by its integration into IBM's Qiskit software development kit, a toolkit used by more than half a million users globally. This significant move was further highlighted at IBM's 2022 annual quantum summit, showcasing the company's commitment to this innovative approach. Moreover, IBM has announced plans to develop its future quantum infrastructure around circuit cutting, signaling a major shift in the landscape of QC. The technology's potential and versatility have also attracted the attention of multiple companies, all actively exploring various use cases to leverage its capabilities. This widespread interest is further validated by the numerous grants and awards received, indicating a strong confidence in the practical applications and future prospects of circuit cutting in the QC industry.

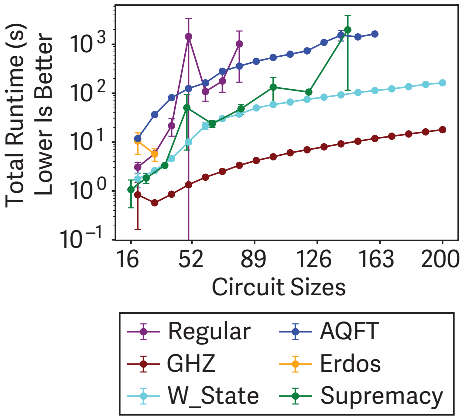


FIGURE 4. End-to-end wall clock runtimes, including cut searching, QPU runtime, and classical postprocessing. Each data point is the average of three trials. The error bars represent the standard deviations. Experiments terminate benchmarks when the estimated tensor contraction exceeds 10^{15} flops. AQFT: approximate quantum Fourier transform; GHZ: Greenberger–Horne–Zeilinger; Regular: 3-regular graphs; Erdos: Erdos–Renyi graphs; Supremacy: Google Quantum Supremacy experiment.

THE FUTURE OF DISTRIBUTED HYBRID COMPUTING

QC circuit cutting makes possible a novel hybrid CPU–QPU computing paradigm, fostering multidisciplinary collaborations and driving real-world applications far beyond mere proof of concept. The growing industry acceptance of these works underscores distributed hybrid computing's emergence as a pivotal aspect of QC. Looking ahead, there are exciting multidisciplinary opportunities to advance practical distributed hybrid computing. They include the following:

1. *Integrating classical high-performance computing techniques*: Bridging the gap between current quantum workloads and state-of-the-art distributed QC requires advancements in tensor network computing and the use of parallel GPUs, application-specified integrated circuits,

and field-programmable gate arrays. For example, practical benchmarks may require between 10^{15} to 10^{20} flops of classical post-processing. As a comparison, GPT-3 training requires about 10^{23} flops.

2. *Designing future hybrid QPU–CPU computing data centers*: The advent of cloud computing data centers for QC opens new avenues in distributed systems, such as optimizing for reduced latency and increased throughput. This involves tackling challenges in load balancing and resource allocation.
3. *Co-designing application and distributed hybrid CPU–QPU computing back ends*: Recognizing distributed QC as the standard back end for running workloads, domain experts across various fields are well positioned to develop more sophisticated and efficient algorithms, specifically optimized for these advanced computing platforms.
4. *Analyzing hybrid QPU–CPU computing complexity*: The future development of hybrid computing relies on a comprehensive theoretical understanding of its advantages and limitations beyond empirical evidence. Theory researchers should take the charge to study the complexities of hybrid systems and guide the development of efficient systems and applications.
5. *Addressing hybrid data security challenges*: Hybrid computing requires communications between quantum and classical back ends and hence may be susceptible to new data leakage channels. Hybrid data security will be the key to enabling trustworthy distributed hybrid computing.

In conclusion, distributed hybrid QPU–CPU computing represents a transformative path forward for QC, bridging the gap between current quantum hardware capabilities and the demands of advanced quantum algorithms. Circuit cutting represents the key methodology for making these distributed hybrid approaches possible. By enabling the execution of large quantum circuits on smaller scale QPUs, this technique not only makes quantum computations more feasible but also expands the range of solvable problems. The integration of classical computing and QC through this method underscores a pivotal shift toward practical, scalable, and efficient QC solutions, setting the stage for a future where complex quantum tasks are more accessible.

ACKNOWLEDGMENT

This article is based upon work supported by the U.S. Department of Energy, Office of Science, National Quantum Information Science Research Centers, and Co-design Center for Quantum Advantage (C2QA) under contract DE-SC0012704. This work is funded in part by EPiQC, a National Science Foundation (NSF) Expedition in Computing, under award CCF-1730449. This article is based upon work supported by (while Martonosi was serving at) the NSF.

REFERENCES

1. I. L. Markov and Y. Shi, “Simulating quantum computation by contracting tensor networks,” *SIAM J. Comput.*, vol. 38, no. 3, pp. 963–981, 2008, doi: 10.1137/050644756.
2. T. Peng, A. W. Harrow, M. Ozols, and X. Wu, “Simulating large quantum circuits on a small quantum computer,” *Phys. Rev. Lett.*, vol. 125, no. 15, 2020, Art. no. 150504, doi: 10.1103/PhysRevLett.125.150504.
3. W. Tang and M. Martonosi, “ScaleQC: A scalable framework for hybrid computation on quantum and classical processors,” 2022, *arXiv:2207.00933*.
4. W. Tang, T. Tomesh, M. Suchara, J. Larson, and M. Martonosi, “CutQC: Using small quantum computers for large quantum circuit evaluations,” in *Proc. 26th ACM Int. Conf. Archit. Support Program. Lang. Oper. Syst.*, 2021, pp. 473–486, doi: 10.1145/3445814.3446758.
5. W. K. Wootters and W. H. Zurek, “A single quantum cannot be cloned,” *Nature*, vol. 299, no. 5886, pp. 802–803, 1982, doi: 10.1038/299802a0.

WEI TANG is a final-year computer science Ph.D. student in Prof. Margaret Martonosis group at Princeton University, Princeton, NJ 08540 USA. Contact him at weit@princeton.edu.

MARGARET MARTONOSI is the H.T. Adams ’35 Professor of Computer Science at Princeton University, Princeton, NJ 08540 USA. She is a Fellow of IEEE and the Association for Computing Machinery. Contact her at mrm@cs.princeton.edu.

EDITOR: Stephen J. Andriole, steve@andriole.com

This article originally appeared in IT Professional vol. 25, no. 1, 2023

COLUMN: LIFE IN THE C-SUITE

New Ways of Working Are Already Old

Stephen J. Andriole, *Villanova School of Business, Villanova, PA, 19085, USA*

The phrase "ways of working" means different things to different industries and, of course, the executives within those industries. Let's try to level set with the help of Lauren Soucy[1] (the bold italics are mine):

> "'New ways of working' was a radical departure from the office-bound corporate culture of the early 20th century. It started with concepts like mobile offices, flexible workspaces, videoconferencing, etc. with the aim to reduce people's burdens by using digitization for automating work processes. Along the same lines, ***today's new ways of working intend to enhance employee experience and employee retention in the workplace***."

The assumption is that new ways of working will keep employees happy, and happy employees will enable greater productivity and, ultimately, profits. "Varied work agreements," cultures of equality, work flexibility, and work/life balance are all part of the process, and it's like nothing we've ever seen.

Let's acknowledge that new ways of working are also now "required." There are changes in the workforce impacting the way companies define traditional roles, processes, structures, protocols and governance. Many of these changes are almost nonnegotiable with Generation Z, where expectations about "work" are fundamentally changing. "The great resignation" and "quiet quitting" describe some of the new relationships with "work" that many professionals embrace. C-suites must understand these relationships if they want to recruit and retain the best talent without alienating longstanding talent that's still valuable:[1]

> "Shifting from a traditional setup to a flexible work environment is a radical change for any company. Older employees and those who're used to the conventional working practices can find it harder to adjust to the new ways . . . companies must strike a balance between attracting new talent and retaining the loyalty of existing employees."

All of that said—and thanks to a global pandemic—the future of work arrived faster than expected. The education and training industries, for example, learned that professors, trainers, and students can live anywhere and that educational and training content can live in the cloud. Health care now relies upon telemedicine, and shopping of all kinds has been permanently impacted by e-commerce. The pandemic also changed management styles. "Back in the day," managers looked over the shoulders of their employees—or just walked down the hall—to check on project progress. However, since the pandemic, managers now often manage by outcome: if project deliverables are on time and of high quality, process management has largely disappeared in many industries. Meetings are now only remote in many companies, while others struggle with hybrid meeting protocols. Yes, some companies are "requiring" employees to return to the office, but some of these companies are losing their best employees who still want to work from home.

Note that assumptions about the permanency of these new ways of working are sometimes challenged by those who want everyone to return to the office to work the way they always did and for managers to, once again, just walk down the hall to check on a project. However, make no mistake—the research suggests that these new ways of working are here to stay.[2,3,4] "Remote," "tele-," and "hybrid" are the new watchwords.

So what should C-suites do?

UNDERSTAND AND COMMIT

There are several levels of understanding here. The first is how we understand abstractions like, for example, global warming. The second is how we understand the real-world implications of global warming, and the third is understanding what we should do about global warming. This last understanding is how executives should understand ways of working. The "future of work"—another phrase made popular by technologists and management consultants—also has levels of understanding. The future of work is the umbrella

1520-9202 © 2023 IEEE
Digital Object Identifier 10.1109/MITP.2023.3236495
Date of current version 10 March 2023.

under which "ways of working" lives. Both concepts can be understood as abstractions, influencers, and agendas. It's the agenda part—"What should I do about it?"—that's the most important now and forever.

Executives must commit to the urgency of how ways of working will impact their profitability, which is the most difficult part of the understanding process because most executives are comfortable with abstract understandings of most things. Why? Because abstract understanding requires acknowledgement ("Yes, we get it"), but no action. C-suiters often prefer to defer decisions because decisions come with accountability. If executives are unwilling to fully commit to new ways of working, then potential benefits will be lost. Therefore, the first test is an executive's agenda-driven commitment to new ways of working.

There are no real choices here. The requirements of the modern professional environment dictate commitment and action despite how some C-suiters might feel about the specific "ways" or their commitment to older ways of working. There's a generational shift already well underway. The very definition of business is changing as new professionals enter the workforce. Turning back the clock is not an option. C-suites that get out in front of these inevitable trends will fare better than those who fight them.

ASSESS CONTEXT AND WAYS

Once a full commitment is made, the C-suite can focus on how it wants to exploit new ways of working. Unlike traditional approaches to leadership, which lie at the heart of all action, the definition and exploitation of ways of working require different approaches. Leadership advice at the abstract level usually includes things like "Listen," "Be empathetic," "Tell the truth," "Be present," "Be consistent"—you get the idea. There are also different leadership styles, like activist leaders, innovative leaders, passivist leaders, democratic leaders, visionary leaders, mentor leaders, and dictatorial leaders.

While abstract advice and cleverly named leadership styles can be useful, leadership around ways of working should be contextual. There are specific factors that will influence both the definition of ways of working and implementation plans. For example, new ways of working are different depending upon the following:

- type of industry;
- status of the company (private, public, not-for-profit, etc.);
- type of business offerings (product, service, etc.);
- composition and activism of stakeholders;
- number and effectiveness of competitors;
- regulatory controls;
- nature and frequency of crises;
- stage of the business (start-up, early stage, etc.);
- digital maturity;
- national, global, or both;
- condition of the business (making revenue projections, etc.);
- definitions of success;
- organizational structure;
- number of "tenured" employees;
- expected outcomes; and so on.

The advice is to avoid generalities at all costs and develop new ways-of-working plans within your specific industry. If you inspect this list of contextual factors, you can locate your industry, your company, and the state of your company, which will determine how you identify and implement the best ways of working for your company.

What are the major ways of working you must assess in your specific context? They include the following:

- flexible work schedules;
- remote work;
- process- versus outcome-based management;
- self-management;
- mobile work;
- virtual collaboration;
- workweek management;
- automation;
- wellness investment; and
- investments in support technology.

As Figure 1 suggests, there's a matrix here that's big and complex.

C-suiters should locate their companies contextually and then visit each of the new ways-of-working options. While not all of the cells require an entry, the matrix should help executives think about how they should transition to new ways of working. For example, if a company is continuously in crisis mode, it may be necessary to rethink its approach to remote work. Companies that sell services versus products might rethink how they define virtual collaboration. Context influences which ways of working are prioritized.

At Tesla, Elon Musk decided to mandate the end of remote work:[5]

> "Everyone at Tesla is required to spend a minimum of 40 hours in the office per week. Moreover, the office must be where your actual colleagues are located, not some remote pseudo-office. If you don't show up, we will assume you have resigned."

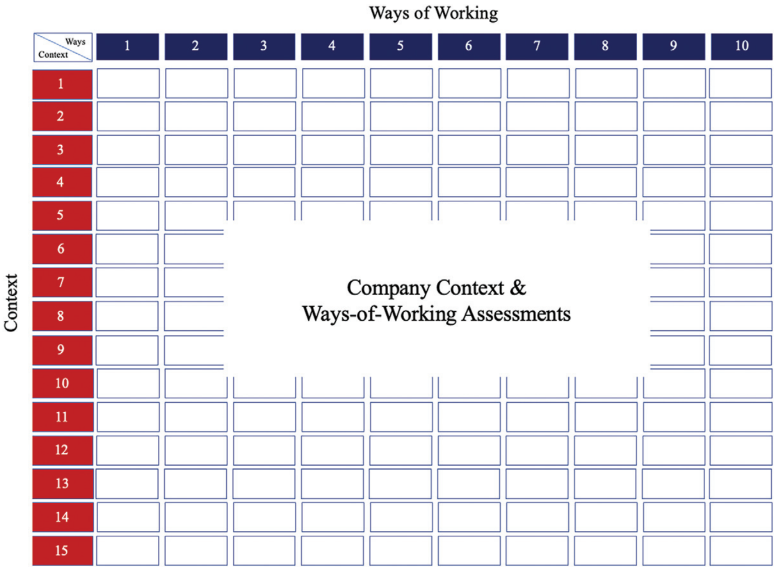


FIGURE 1. Context and ways.

Are mandates effective? Or is it better to explain why some new ways of working make sense for the company and why others don't? C-suites should explain which ways of working they plan to adopt and which they will avoid—for now or permanently. The matrix in Figure 1 can provide some guidance to this process.

FINAL THOUGHTS

Ten years ago, no one would have predicted that higher education, complex training, and medical checkups would be routinely delivered remotely (just as no one would have predicted a global pandemic). "Tele-" is the favored approach to lots of problem solving. The adoption of new ways of working is inevitable. Generational changes are accelerating adoption. Fortunately, digital technology is available that enables most of the new ways of working—technology that's growing in capability and use: who would ever have predicted the ease with which grandparents would Zoom their kids and grandkids during the pandemic? C-suites can recruit the best employees—and succeed—by adopting new ways of working. The C-suites that fail to adopt new ways of working will continue to live comfortably in the 20th century.

REFERENCES

1. L. Soucy, "New ways of working for modern organizations." Time Doctor. [Online]. Available: https://www.timedoctor.com/blog/new-ways-of-working/
2. B. Robinson, "Remote work is here to stay and will increase into 2023, experts say," *Forbes*, Feb. 2022. [Online]. Available: https://www.forbes.com/sites/bryanrobinson/2022/02/01/remote-work-is-here-to-stay-and-will-increase-into-2023-experts-say/?sh=263797e220a6
3. P. Choudhury, "Our work-from-anywhere future best practices for all-remote organizations," *Harvard Bus. Rev.*, 2020. [Online]. Available: https://hbr.org/2020/11/our-work-from-anywhere-future
4. D. Gerdeman, "COVID killed the traditional workplace. What should companies do now?" Harvard Bus. School Work. Knowl., Boston, MA, USA, 2021. [Online]. Available: https://hbswk.hbs.edu/item/covid-killed-the-traditional-workplace-what-should-companies-do-now
5. K. Walters, "Elon Musk has a clear work-from-home policy," *Accounting Today*, Jun. 2022. [Online]. Available: https://www.accountingtoday.com/opinion/elon-musk-has-a-clear-work-from-home-policy

STEPHEN J. ANDRIOLE is the Thomas G. Labrecque Professor of Business Technology with the Villanova School of Business, Villanova, PA, 19085, USA, where he researches and teaches in the emerging technology, artificial intelligence, and machine learning areas. He is the former director of cybernetics technology at the Defense Advanced Research Projects Agency and the chief technology officer at Cigna Corporation and Safeguard Scientifics. He was a professor of information systems and electrical and computer engineering at Drexel University; at George Mason University, he was the George Mason Institute Professor Information Technology and the chair of the Department of Information Systems and Software Engineering. Contact him at stephen.andriole@villanova.edu and at https://andriole.com.

EDITOR: Robert Blumen, Katana Graph, robert@robertblumen.com

This article originally appeared in IEEE Software vol. 40, no. 3, 2023

DEPARTMENT:
SOFTWARE ENGINEERING RADIO

Jon Smart on Patterns and Antipatterns for Enterprise Software Success

Brijesh Ammanath

FROM THE EDITOR

In Episode 543 of "Software Engineering Radio," Jon Smart, business agility practitioner, thought leader, coach, and lead author of *Sooner Safer Happier: Patterns and Antipatterns for Business Agility*, discusses patterns and antipatterns for the success of enterprise software projects with host Brijesh Ammanath. Topics include patterns and principles needed in the digital age; why doing an agile, lean, or DevOps transformation is an antipattern; outcomes versus outputs; and the importance of psychological safety, mindset changes, and transformational leadership. We provide summary excerpts below; to hear the full interview, visit http://www.se-radio.net or access our archives via RSS at http://feeds.feedburner.com/se-radio. —*Robert Blumen*

Brijesh Ammanath: What is business agility, and why is it important?

Jon Smart: It is improving ways of working to improve outcomes. The first industrial revolution went from craft working to division of labor and working in a factory. Today, you can trace the DNA of ways of working in large organizations back to 1771 and the very first factories. Business agility is triggered by the latest technologically led revolution, which is the age of digital, and it's driven by competitive pressure—no longer can companies take a long time to have a feedback loop on the value they're producing.

For companies of all sizes, it doesn't take long to end up with the same level of dysfunction; people start to introduce more process and more gated controls and to slow down the flow of value. It's a human characteristic to keep adding more process and bureaucracy.

Digital Object Identifier 10.1109/MS.2023.3248899
Date of current version: 18 April 2023

How are outcomes different from outputs?

Outputs work in a factory-type scenario. In the age of oil and mass production, the focus was on output, and the definition of productivity was the number of units of output per unit of input. An output is a thing, deliverable, widget, piece of software, or system. The common antipattern here at organizations is a relentless focus on output, with hardly any focus on the outcome.

The reason for the output is to achieve a certain outcome. Outcome might be increased market share, increased revenue, support for more first-time buyers, increased market share in Latin America. The output is how you might achieve the outcome. This is a big mindset shift that includes an experimentation mindset. Because change is unique and unknowable, and the only way to learn is by doing, we have to run experiments. And we have to minimize time to learning and feedback so that we can pivot and have the cheapest cost of failure. Milestones in a project plan don't define success. The definition of success is the key results in an OKR: objectives and key results.

2469-7087/25 © 2025 IEEE

SOFTWARE ENGINEERING RADIO

Visit www.se-radio.net to listen to these and other insightful hour-long podcasts.

RECENT EPISODES

- 547—Nicholas Manson, a software-as-a-service architect with more than two decades of experience building cloud applications, speaks with host Kanchan Shringi about identity and access management requirements for cloud applications.
- 544—Ganesh Datta, chief technology officer and cofounder of Cortex, joins Priyanka Raghavan to discuss site reliability engineering versus DevOps.
- 538—Roberto Di Cosmo, professor of computer science at University Paris Diderot and founder of the Software Heritage Initiative, discusses with Gavin Henry the reasons for and challenges of the long-term archiving of publicly available software.
- 535—Dan Lorenc, chief executive officer of Chainguard, a software supply chain security company, joins "Software Engineering Radio" editor Robert Blumen to talk about software supply chain attacks.

UPCOMING EPISODES

- Luca Casonato talks with host Jeremy Jung about Deno and Deno Deploy.
- Matt Frisbee, author of *Building Browser Extensions*: *Create Modern Extensions for Chrome, Safari, Firefox, and Edge*, speaks with Kanchan Shringi about browser extensions, including key areas where theyve been successful
- Alex Boten, author of *Cloud Native Observability With OpenTelemetry*: *Learn to Gain Visibility Into Systems by Combing Training, Metrics, and Logging With OpenTelemetry*, joins host Robert Blumen for a conversation about software telemetry and the OpenTelemetry project.

What is psychological safety, and why is it important?

It's the ability to feel safe, ask questions, challenge authority, have your voice heard, express your thoughts without fear of repercussion or of being shut down or belittled. It's the ability to have open vulnerable conversations with respect. It's also about not having a blame culture. If something goes wrong, it's not because somebody did something wrong. It's because there was something in the system of work that enabled this thing to happen.

What is system entropy, and how is it related to technical excellence?

Human and technical systems go backward over time and get worse. With technology, there's more technical debt, more treacle, and it takes longer to get stuff done. In the case of human systems, there's more bureaucracy, more process, more approvals and committees. We humans like to keep adding processes, rather than taking things away.

Given that human and technical systems have entropy over time, it is necessary to continuously improve in terms of culture, process, and technology. Technical excellence is important here because if the focus is just on agile or Scrum, there's nothing there about technical excellence. So, it's important to have a focus on technical excellence continuously, to continuously refactor and improve, and to dedicate some bandwidth to that continuous improvement. Improving daily work is as important as daily work, and not doing that, you end up going slow to go faster, or you will end up going slower. What that means is make sure to put time aside to improve—in the case of technology for refactoring—and continuously improve not only the technology but also your ways of working.

How do you measure technical excellence?

It all comes back to the lagging measures of time to value. So, lead time, flow efficiency, the amount of time that work has been worked on versus the end-to-end time, quality—and by quality, I mean failure demand, not defects; I mean unplanned work, which is failure demand. That's another measure, quality, and safety, so in terms of better value: sooner, safer, happier.

The words *safety* and *safer*, in particular, have to do with mandated controls, like information security, data privacy, encrypting data at move, encrypting data at rest, two-factor authentication, not making the newspaper headlines because customer data has been hacked and leaked onto the Internet.

And then obviously, *happier*. That's happier colleagues as much as it is happier customers. That's something that's missing from the DevOps Research and Assessment metrics from "Accelerate"—the word *happier* is not there. That's the one that's usually missed. So, time to value, lead time, and flow efficiency, quality, safety, and happiness, all of which lead to business value.

How are software development methodologies evolving? What will be the new ways of working?

There's still a big gap between business and technology. There's a lot of room for improvement in terms of properly multidisciplinary teams, which are business and IT together. We can still have reporting lines into technology and into nontechnology, but work is not going to go through the reporting line. The reporting line is there for personal development, career growth, and care for the individual. Work is flowing in the value stream.

It is increasingly going to be BizDevOps, which for me is a bit back to the future. This is how we used to work in the early 1990s. This breaking down the barriers, even for some Silicon Valley companies, even with the notion of product and engineering, there still ends up being a business silo, a product silo, and an engineering silo. And the product silo is just doing what business analysts used to do in the past, which is talking to the business, writing requirements, and handing them to engineering, which is not great. We need more outcome-focused value stream alignment — BizDevOps.

BRIJESH AMMANATH, is a volunteer host at SERadio, Pune 411040, India. Contact him at akbrijesh@gmail.com.

Unlock Your Potential

WORLD-CLASS CONFERENCES — Over 195 globally recognized conferences.

DIGITAL LIBRARY — Over 900k articles covering world-class peer-reviewed content.

CALLS FOR PAPERS — Write and present your ground-breaking accomplishments.

EDUCATION — Strengthen your resume with the IEEE Computer Society Course Catalog.

ADVANCE YOUR CAREER — Search new positions in the IEEE Computer Society Jobs Board.

NETWORK — Make connections in local Region, Section, and Chapter activities.

Explore membership today at the IEEE Computer Society
www.computer.org

IEEE COMPUTER SOCIETY
IEEE

IEEE COMPUTER SOCIETY D&I FUND

Drive Diversity & Inclusion in Computing

Supporting projects and programs that positively impact diversity, equity, and inclusion throughout the computing community.

DONATE TODAY!

IEEE COMPUTER SOCIETY

IEEE Foundation

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

This article originally appeared in
Computer
vol. 57, no. 10, 2024

DEPARTMENT: GAMES

Integrating Blockchain Technology in Online Gaming Ecosystems

Kevin Macwan, *Amazon*

Future research and collaboration between blockchain and game developers are essential to address challenges such as scalability, regulatory compliance, and system integration and fully realize blockchain's potential in creating secure, transparent, and engaging gaming experiences.

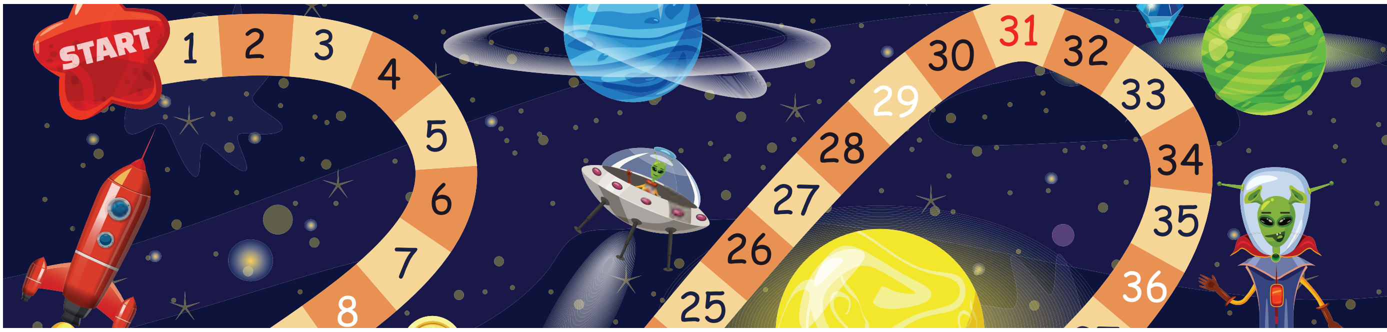
Blockchain technology, characterized by its decentralized, immutable, and transparent nature, holds transformative potential across various industries, including online gaming.[1] Blockchain operates through a distributed ledger that records transactions across multiple computers, ensuring that the recorded information is secure, transparent, and resistant to tampering.[2] This technology employs cryptographic techniques to create and verify transactions to provide a high level of security and trust without the need for intermediaries. Smart contracts, which are self-executing contracts with the terms directly written into code, further enhance blockchain's utility by automating processes and reducing the need for manual intervention.[3] The transformative potential of blockchain in online gaming lies in its ability to provide secure, transparent, and decentralized solutions for in-game transactions, asset ownership, and player interactions. According to Vision Research Reports[4] As shown in Figure 1, the market potential of Blockchain-based gaming is expected to reach US$887 billion in 2030.

The current state of online gaming ecosystems is characterized by centralized servers and databases managed by game developers and publishers.[5] These traditional systems often face challenges related to security, transparency, and trust.[6] Issues such as hacking, fraud, and lack of actual ownership of in-game assets are dominant.[7] Furthermore, centralized systems mean that players are dependent on game developers for the integrity and availability of their gaming assets and experiences.[8] Blockchain technology offers a solution to these challenges by decentralizing the control and management of game assets and transactions.[9] It enables true ownership of digital assets through nonfungible tokens (NFTs), enhances data security through its immutable ledger, and promotes transparency in in-game economies. By utilizing blockchain, the online gaming ecosystem can achieve a higher level of security, fairness, and innovation, driving the industry toward a more decentralized and player-centric future.

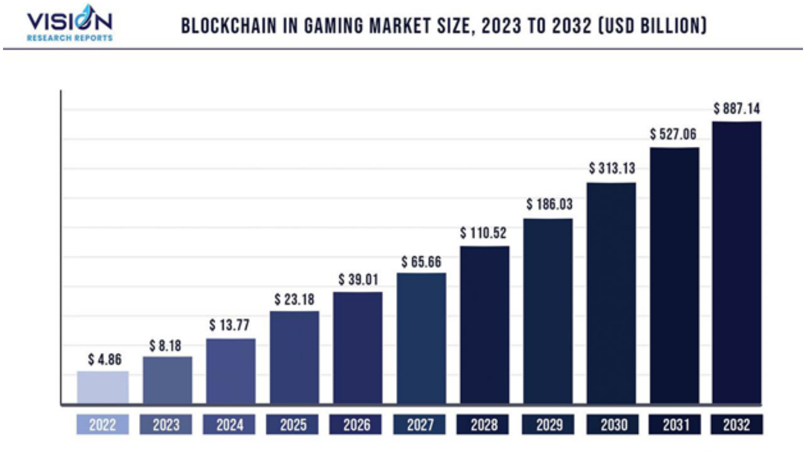
The primary objective of this article is to investigate the integration of blockchain technology within online gaming ecosystems, focusing on how blockchain can address existing challenges and introduce new opportunities. This involves a detailed examination of the technical mechanisms by which blockchain can be incorporated into game development, in-game transactions, and asset management. Additionally, this study aims to identify and discuss the potential benefits and challenges associated with such integration. Possible benefits include enhanced security for in-game transactions, true digital ownership for players, and the creation of decentralized gaming platforms that empower users. Conversely, the challenges consist of technical hurdles such as scalability and interoperability, economic considerations like the impact on existing game economies, and regulatory issues pertaining

Digital Object Identifier 10.1109/MC.2024.3431908
Date of current version: 23 September 2024

2469-7087/25 © 2025 IEEE



to data privacy and compliance. By exploring these aspects, the study seeks to provide a comprehensive understanding of how blockchain technology can revolutionize online gaming. Ultimately, it will offer a roadmap for developers, researchers, and stakeholders interested in utilizing blockchain to enhance the gaming experience.



Source: www.visionresearchreports.com

FIGURE 1. Blockchain gaming market size according to Vision Research Reports.[4]

TECHNOLOGICAL BACKGROUND

Blockchain fundamentals

Blockchain technology is fundamentally a decentralized, distributed ledger system that records transactions across multiple computers, ensuring that the recorded information is immutable.[10] Each block in a blockchain contains a list of transactions, and these blocks are linked together in chronological order, forming a chain. A critical aspect of blockchain is its use of consensus mechanisms, such as proof of work (PoW) or proof of stake (PoS), to validate and record transactions without the need for central authority. Smart contracts are self-executing contracts with the terms of the agreement directly written into code.[11] These contracts automatically enforce and execute the terms of the agreement when predefined conditions are met, providing automation and reducing the need for intermediaries. This combination of distributed ledgers, consensus mechanisms, and smart contracts creates a robust and secure framework that can revolutionize various industries, including gaming.[7]

Gaming ecosystems

Current online gaming ecosystems are typically centralized, with game developers and publishers maintaining control over the servers and databases that host game data and manage in-game transactions.[12] These ecosystems rely heavily on traditional client-server architectures, where the game client interacts with a central server that processes and verifies all game-related activities. This centralized approach often leads to several issues, such as vulnerability to hacking, fraud, and server downtimes, which can disrupt the gaming experience.[13] Moreover, players need to have true ownership of their in-game assets, as these assets are stored and controlled by the game developers. Technological frameworks within these ecosystems include various programming languages, game engines (such as Unity and Unreal Engine), and networking protocols that facilitate multiplayer interactions and online gameplay. Despite advancements in these technologies, the centralized nature of current gaming ecosystems poses significant limitations in terms of security, transparency, and player empowerment (Figure 2).

Intersection of blockchain and gaming

The integration of blockchain technology into online gaming ecosystems presents numerous potential intersection points that can address the limitations of centralized systems.[12] One of the primary integration

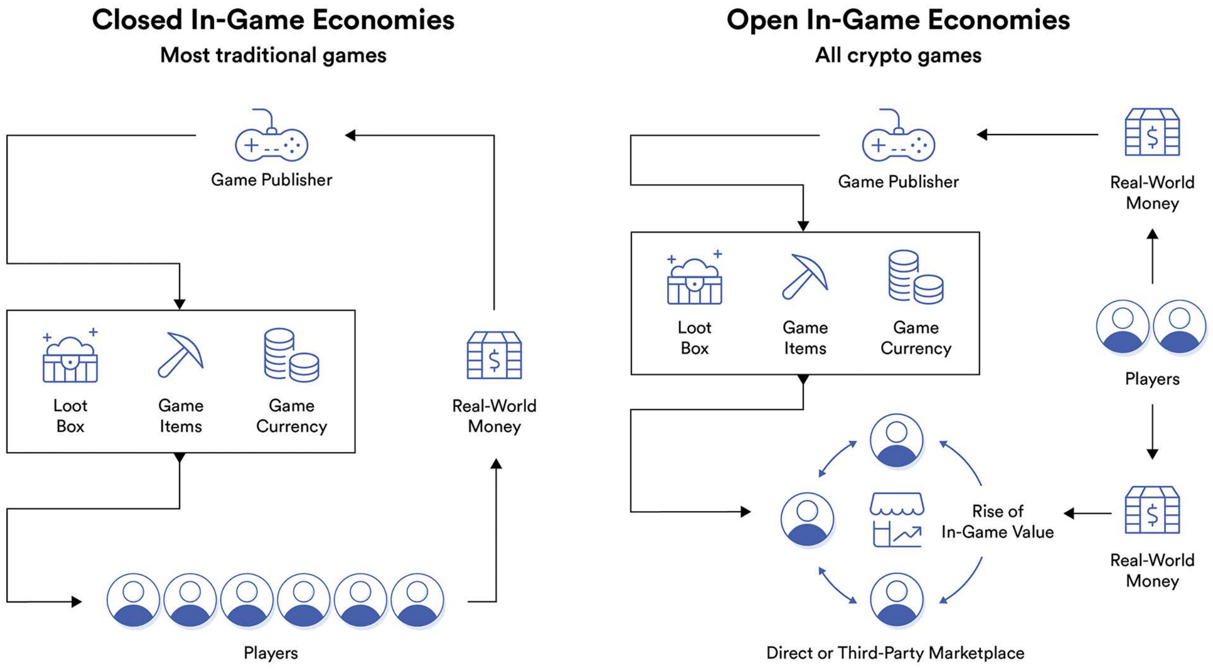


FIGURE 2. A comparison of traditional gaming economics versus blockchain-based gaming economics.[14]

points is the use of blockchain for secure and transparent in-game transactions. By utilizing blockchain's distributed ledger, transactions can be recorded immutably, reducing the risk of fraud and ensuring transparency. Another critical integration point is the use of NFTs to represent in-game assets, allowing players to have true ownership and the ability to trade these assets across different platforms. Smart contracts can automate various aspects of game mechanics and in-game economies, ensuring fairness and reducing the need for manual oversight.[15] Additionally, blockchain can facilitate the development of decentralized gaming platforms, where the control and governance of the game are distributed among the players rather than being centralized with the developers. This decentralization can lead to more resilient and player-centric gaming experiences, empowering users and fostering innovation within the gaming industry.

SECURITY ENHANCEMENTS THROUGH BLOCKCHAIN

Data integrity and transparency

Blockchain technology ensures data integrity in gaming by utilizing its decentralized ledger system, where each transaction or piece of data are cryptographically linked to the previous one in a chain of blocks.[16] This structure makes it virtually impossible to alter any part of the chain without affecting all subsequent blocks, thereby ensuring that data remains consistent and tamper-proof. In the context of gaming, this immutability guarantees that player actions, transactions, and in-game assets are recorded accurately and cannot be retroactively modified or deleted. Furthermore, the transparency inherent in blockchain systems allows all participants to view the history of transactions, fostering trust among players and developers. This transparency is critical in online gaming environments where trust is paramount, as it provides an auditable trail of all in-game activities, thereby reducing disputes and enhancing the overall gaming experience.

Anticheating mechanisms

The implementation of anticheating mechanisms using blockchain technology addresses one of the most persistent issues in online gaming.[17] Traditional anticheating measures rely on centralized servers to detect and prevent cheating behaviors, which sophisticated attackers can bypass. Blockchain, however, offers a decentralized approach where game logic and player interactions can be encoded into smart contracts that are executed deterministically and transparently. These smart contracts can validate player actions against predefined rules, ensuring that any attempt to cheat is automatically detected and invalidated. For example, in multiplayer games, the outcomes of actions (like damage calculations in a battle) can be verified by the blockchain network, making it difficult for players to alter game data for unfair advantages. Additionally, the distributed nature of blockchain means that there is no single point of failure or target for hackers, significantly increasing the difficulty of executing widespread cheating schemes.

Fraud prevention

Blockchain enhances security in gaming transactions and in-game assets by providing a robust framework for fraud prevention.[12] Each transaction recorded on the blockchain is timestamped and linked to previous transactions, creating a traceable and unalterable history. This traceability is particularly beneficial in preventing fraud in the trading of in-game assets, where players can buy, sell, and trade items with confidence. The use of NFTs ensures that each in-game asset is unique and owned by the player, with ownership verifiable on the blockchain. This eliminates the risk of counterfeit items and ensures that players genuinely own their digital assets. Additionally, blockchain's consensus mechanisms prevent double-spending, where a player could try to use the same in-game currency or asset multiple times.[18] By distributing the verification process across multiple nodes, the blockchain ensures that each transaction is valid and prevents fraudulent activities from compromising the gaming ecosystem.

DECENTRALIZATION AND OWNERSHIP

True ownership of in-game assets

The utilization of NFTs in gaming represents a significant shift toward true ownership of in-game assets.[13] NFTs are unique digital tokens verified on the blockchain, which can represent a wide range of digital items such as weapons, characters, or virtual real estate. Unlike traditional gaming assets that are stored on centralized servers and ultimately controlled by game developers, NFTs are stored on a decentralized blockchain, ensuring that the player has full ownership and control over their assets.[19] This ownership is immutable and transferable, meaning players can buy, sell, or trade their assets on various platforms without relying on the game developer's infrastructure. The decentralized nature of NFTs ensures that even if a game server shuts down, the player's assets remain intact and accessible, providing a level of security and permanence previously unavailable in the gaming industry.

Interoperability of assets

Blockchain technology facilitates the interoperability of assets across different games and platforms, a concept that significantly enhances the gaming experience.[20] By standardizing the representation and ownership of in-game assets through NFTs, players can seamlessly transfer their items from one game to another. This interoperability is achieved through smart contracts and blockchain protocols that define and enforce the rules for asset transfer and utilization. For instance, a sword acquired in one fantasy game could be used in another, or a piece of virtual real estate in a metaverse could serve as a hub for various games. This cross-game asset utilization not only increases the value and utility of in-game items but also fosters a more interconnected and expansive gaming ecosystem. Developers can collaborate to create shared economies and ecosystems, enhancing player engagement and offering new revenue streams through cross-platform collaborations.

Decentralized game development and governance

Decentralized game development and governance represent a fundamental transformation in how games are created and managed. Traditional game development is centralized, with decisions made by a core team of developers and publishers. In contrast, decentralized game development leverages blockchain technology to distribute decision-making processes across the community of players and developers.[1] Through mechanisms such as decentralized autonomous organizations (DAOs), game stakeholders can participate in voting on game updates, features, and governance policies. This community-driven approach ensures that the game's development aligns more closely with the players' interests and desires,

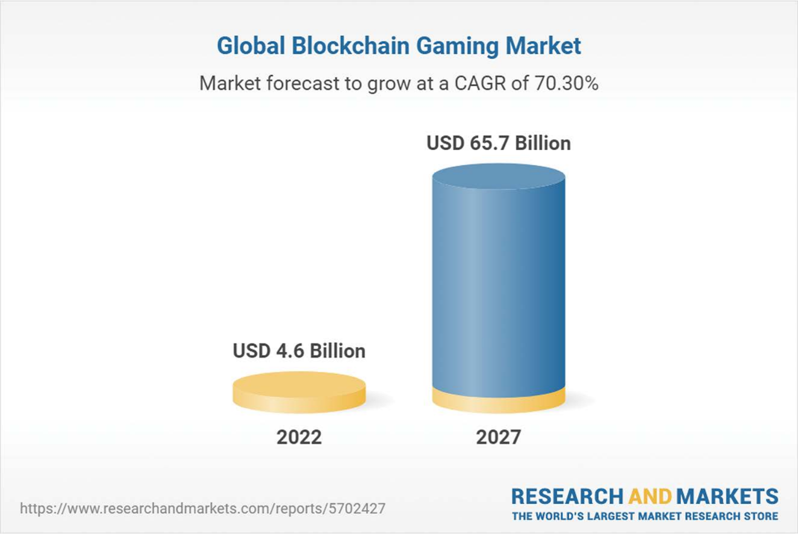


FIGURE 3. Economic and market implications of blockchain gaming ecosystem.[22]

fostering a more inclusive and democratic gaming environment. Additionally, decentralized governance can enhance transparency and trust, as all decisions and transactions are recorded on the blockchain and accessible to all participants. This model not only empowers players but also creates a more resilient and adaptive game development process, capable of responding more effectively to community feedback and emerging trends.

ECONOMIC AND MARKET IMPLICATIONS

The integration of blockchain technology profoundly impacts in-game economies by introducing new mechanisms for value creation and exchange.[21] Blockchain enables the creation of decentralized in-game currencies and assets, which are verifiable and transferable on the blockchain. This verifiability ensures that in-game assets are not duplicable, enhancing their value and rarity. The decentralized nature of blockchain also allows for peer-to-peer transactions without intermediaries, reducing transaction costs and increasing the fluidity of in-game markets. Players can trade assets securely and transparently, which can lead to more dynamic and robust in-game economies. Moreover, the use of smart contracts automates and secures complex economic interactions, such as auctions, lending, and staking of assets, fostering an environment where economic activities can flourish with minimal friction. According to Research and Markets,[22] The global blockchain gaming market is projected to grow from US$4.6 billion in 2022 to US$65.7 billion by 2027 at a compound annual growth rate of 70.3% during the forecast period. Rising funding for blockchain games is one factor driving the market growth (Figure 3).

Blockchain technology paves the way for innovative business models and revenue streams within the gaming industry. One notable development is the play-to-earn model, where players are rewarded with cryptocurrency or NFTs for their participation and achievements in games.[23] This model incentivizes engagement and provides players with tangible value for their time and skills. Additionally, blockchain enables fractional ownership and investment in virtual assets, allowing players to own and benefit from high-value items or properties collectively. Developers can also explore new monetization strategies, such as tokenized crowdfunding, where players can invest in game development projects and receive tokens representing a share of the games future profits. This democratizes the funding process and aligns the interests of developers and players, fostering a more collaborative and invested gaming community.

TECHNICAL CHALLENGES AND SOLUTIONS

Scalability issues

Addressing the scalability issues of blockchain in gaming is critical to ensuring that blockchain-based gaming platforms can handle the high transaction volumes and complex interactions characteristic of modern online games.[24] Scalability in blockchain refers to the network's ability to process a high number of transactions per second (TPS) without compromising speed or security. Current blockchain networks, such as Bitcoin and Ethereum, face significant scalability challenges due to their consensus mechanisms. For instance, Bitcoin's PoW and Ethereum's existing implementation limit TPS to low double digits, which is insufficient for high-demand gaming environments where thousands of transactions per second may be necessary. Solutions like layer-2 scaling (for example, state channels, sidechains) and next-generation blockchains (for example, Ethereum 2.0's PoS and sharding, Solana's Proof of History) are being developed to address these limitations. Layer-2 solutions offload transactions from the main blockchain, reducing congestion and increasing throughput, while sharding divides the blockchain into smaller, more manageable pieces (shards), each capable of processing its transactions in parallel, thus significantly enhancing overall network capacity.

Integration with existing systems

Integrating blockchain with existing game infrastructures presents several technical challenges due to differences in architecture and operation between traditional gaming platforms and blockchain networks.[25] Traditional game servers are centralized, whereas blockchain operates on a decentralized network, requiring a fundamental shift in how data are managed, and transactions are processed. One significant challenge is ensuring compatibility between the centralized databases of existing games and the decentralized ledgers of blockchain. This often involves creating hybrid systems where certain game functions remain on centralized servers while transactions and asset ownership are managed on the blockchain. Another area for improvement is the latency introduced by blockchain transactions, which can be slower compared to traditional systems due to the time required for consensus mechanisms. This latency can affect the real-time performance essential for many online games. Middleware solutions and application programming interfaces are being developed to bridge these systems, enabling seamless data transfer and synchronization between traditional game servers and blockchain networks. Furthermore, integrating smart contracts to automate in-game transactions and enforce rules requires meticulous programming and thorough security audits to prevent exploits and vulnerabilities.

Performance optimization

Optimizing the performance and efficiency of blockchain-based gaming platforms involves implementing techniques that address the inherent limitations of blockchain technology while enhancing user experience.[26] One approach is the use of efficient consensus algorithms like PoS, which offer faster transaction times and lower energy consumption compared to PoW. Additionally, off-chain transactions and layer-2 solutions, such as the Lightning Network or Plasma, allow for high-speed, low-cost transactions by conducting most operations off the main blockchain and only recording the final state on-chain. Another critical technique is optimizing smart contract execution to reduce gas fees and processing time, which can be achieved by streamlining code and minimizing on-chain computations. Techniques such as state channels enable private, off-chain communication between parties, recording only the outcome on the blockchain, thus reducing the burden on the network. Furthermore, optimizing data storage through distributed storage solutions like InterPlanetary File System (IPFS) ensures that significant game assets do not clog the blockchain, preserving its efficiency. These strategies collectively enhance the scalability, speed, and overall performance of blockchain-based gaming platforms, making them viable for large-scale adoption and use.

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.

Compliance and legal challenges

Navigating the complex field of global and regional regulations presents a significant challenge for integrating blockchain technology into online gaming.[27]

Each jurisdiction has its own set of laws and regulatory frameworks governing the use of blockchain and cryptocurrencies, often leading to a fragmented legal environment. For instance, while some countries have embraced blockchain technology, others impose stringent regulations or outright bans. Compliance involves ensuring that all blockchain transactions and operations within the gaming platform adhere to these diverse regulatory requirements. This includes adhering to anti-money laundering (AML) and know your customer (KYC) regulations to prevent illegal activities such as money laundering and fraud. Additionally, game developers must stay updated with evolving legal standards and ensure their platforms are adaptable to regulatory changes, which can be resource-intensive and complex.

Intellectual property and data privacy

Protecting intellectual property (IP) and ensuring data privacy are critical concerns when integrating blockchain into gaming.[28] Blockchain's transparency and immutability can make it challenging to control the dissemination of proprietary game assets and code, potentially leading to IP infringements. Developers must implement robust IP protection strategies, such as tokenizing IP rights and using smart contracts to enforce usage terms automatically. Data privacy is another significant challenge, as blockchain's immutable nature conflicts with data protection laws like the General Data Protection Regulation (GDPR), which grants individuals the right to be forgotten. Ensuring compliance with data privacy regulations requires innovative solutions, such as off-chain data storage and zero-knowledge proofs, which allow for data validation without exposing the data itself. Balancing the benefits of blockchain transparency with the need for IP protection and data privacy requires meticulous planning and the development of new technical and legal strategies.

PRACTICAL APPLICATIONS AND FUTURE DIRECTIONS

Several successful integrations of blockchain technology into gaming have demonstrated their potential to revolutionize the industry. One notable example is "CryptoKitties," a blockchain-based game that allows players to breed, trade, and sell virtual cats using Ethereum.[29] Each CryptoKitty is a unique NFT, ensuring true ownership and rarity. Another significant case is "Axie Infinity," a play-to-earn game where players collect, breed, and battle creatures called Axies, earning cryptocurrency in the process. The game has created a robust in-game economy driven by NFTs and blockchain-based governance, showcasing how blockchain can enable new economic models in gaming. These examples highlight the transformative potential of blockchain by providing secure ownership, transparency, and innovative economic incentives for players.

From these successful implementations, several key lessons have emerged. First, user experience is vital.[12] While blockchain offers numerous advantages, the complexity of managing digital wallets and understanding blockchain concepts can be a barrier to mainstream adoption. Simplifying the user interface and providing clear instructions can help bridge this gap. Second, scalability remains a critical issue. High transaction fees and network congestion, as experienced by CryptoKitties during its peak popularity, underscore the need for scalable solutions such as layer-2 protocols or more efficient consensus mechanisms. Third, regulatory compliance is essential. Ensuring that blockchain games adhere to local and international regulations, particularly regarding financial transactions and data privacy, is crucial for long-term viability. Finally, community engagement and governance play a vital role. Games like Axie Infinity have shown that involving the community in decision-making processes through decentralized governance models can drive growth and create a loyal player base.

As blockchain technology continues to evolve, several emerging trends and research areas are poised to shape the future of blockchain gaming. One significant trend is the integration of blockchain with augmented reality (AR) and virtual reality (VR) to create immersive and interactive gaming experiences.[30] Research into cross-chain interoperability is also gaining momentum, aiming to enable seamless asset transfer between different blockchain networks, thereby enhancing the utility and liquidity of in-game assets. Another crucial area is the development of more scalable and efficient consensus mechanisms, such as sharding and PoS, to address the current limitations of blockchain scalability and transaction throughput. Furthermore, the rise of DeFi within gaming ecosystems presents new opportunities for creating innovative economic models where players can lend, stake, and earn interest on their digital assets.[31] These trends not only highlight the potential for technological advancements but also underscore the need for continuous research to overcome existing challenges and fully realize the benefits of blockchain in gaming.

Upcoming technological innovations are set to have a profound impact on blockchain gaming, driving both technical and economic transformations. One key innovation is the advancement of zero-knowledge proofs, which can enhance privacy and scalability by allowing transactions to be verified without revealing the underlying data. Another promising technology is the development of decentralized identity solutions, enabling secure and portable digital identities that can be used across multiple gaming platforms. Additionally, the integration of artificial intelligence (AI) with blockchain can lead to the creation of dynamic and adaptive gaming environments where AI-driven NPCs and game mechanics are securely managed and verified on the blockchain.[1] Collaboration opportunities between blockchain developers and game developers are crucial for driving these innovations. Joint initiatives can focus on developing standardized protocols for asset interoperability, creating user-friendly interfaces to simplify blockchain interactions, and exploring new business models that leverage the unique capabilities of blockchain technology. By fostering collaboration and innovation, the gaming industry can harness the full potential of blockchain to create more secure, transparent, and engaging gaming experiences.

In conclusion, the integration of blockchain technology into online gaming ecosystems offers transformative potential by enhancing security, ensuring true ownership of in-game assets, and enabling innovative economic models. However, it also presents significant challenges such as scalability issues, regulatory compliance, and the need for seamless integration with existing systems. Future research and technological advancements, particularly in areas like AR/VR integration, cross-chain interoperability, and decentralized identity solutions, are essential for addressing these challenges and unlocking new opportunities. Collaborative efforts between blockchain developers and game developers will be crucial in driving these innovations and creating more secure, transparent, and engaging gaming experiences. As the gaming industry continues to evolve, blockchain's role will likely become increasingly central, paving the way for a new era of digital interaction and economic potential.

REFERENCES

1. C. Laroiya, D. Saxena, and C. Komalavalli, “Applications of blockchain technology,” in *Handbook of Research on Blockchain Technology*, Amsterdam, The Netherlands: Elsevier, 2020, pp. 213–243.
2. R. Sapra and P. Dhaliwal, “Blockchain: The new era of technology,” in *Proc. 5th Int. Conf. Parallel Distrib. Grid Comput. (PDGC)*, Piscataway, NJ, USA: IEEE Press, 2018, pp. 495–499.
3. S. Nzuva, “Smart contracts implementation, applications, benefits, and limitations,” *J. Inf. Eng. Appl.*, vol. 9, no. 5, pp. 63–75, 2019.
4. “Blockchain in gaming market (by game type: Role playing games, open world games, collectible games; by platform; by device) - Global industry analysis, size, share, growth, trends, revenue, regional outlook and forecast 2023–2032.” Vision Research Reports. Accessed: Jul. 17, 2024. [Online]. Available: https://www.visionresearchreports.com/blockchain-in-gaming-market/40132
5. S. Casper, M. Miozzo, and C. Storz, “The emergence of an entrepreneurial ecosystem: The interplay between early entrepreneurial activity and public policy in the Korean online gaming industry,” *Ind. Innov.*, vol. 31, no. 3, pp. 280–310, Mar. 2024, doi: 10.1080/13662716.2023.2254261.
6. R. McCall and L. Baillie, “Ethics, privacy, and trust in serious games,” in *Handbook of Digital Games and Entertainment Technologies*, R. Nakatsu, M. Rauterberg, and P. Ciancarini, Eds., Singapore: Springer-Verlag, 2017, pp. 611–640.
7. J. T. Holden and S. C. Ehrlich, “Esports, skins betting, and wire fraud vulnerability,” *Gaming Law Rev. Econ. Regulation Compliance Strategy*, vol. 21, no. 8, pp. 566–574, Oct. 2017, doi: 10.1089/glr2.2017.2183.
8. A. Manzoor, M. Samarin, D. Mason, and M. Ylianttila,

"Scavenger hunt: Utilization of blockchain and IoT for a location-based game," *IEEE Access*, vol. 8, pp. 204,863–204,879, 2020, doi: 10.1109/ACCESS.2020.3037182.

9. K. B. Muthe, K. Sharma, and K. E. N. Sri, "A blockchain based decentralized computing and NFT infrastructure for game networks," in *Proc. 2nd Int. Conf. Blockchain Comput. Appl. (BCCA)*, Piscataway, NJ, USA: IEEE Press, 2020, pp. 73–77.
10. G. Habib, S. Sharma, S. Ibrahim, I. Ahmad, S. Qureshi, and M. Ishfaq, "Blockchain technology: Benefits, challenges, applications, and integration of blockchain technology with cloud computing," *Future Internet*, vol. 14, no. 11, 2022, Art. no. 341, doi: 10.3390/fi14110341.
11. J. Madir, "Smart contracts-self-executing contracts of the future?" *Int. House Counsel J*, vol. 13, no. 51, p. 1, 2020.
12. D. Stamatakis, D. G. Kogias, P. Papadopoulos, P. A. Karkazis, and H. C. Leligou, "Blockchain-powered gaming: Bridging entertainment with serious game objectives," *Computers*, vol. 13, no. 1, 2024, Art. no.14, doi: 10.3390/computers13010014.
13. R. Maloul and L. Chevalier, "Study on digital ownership in the gaming industry and analysis of a possible new approach via the implementation of blockchain and non-fungible tokens." DIAL@UCLouvain. Accessed: Jul. 17, 2024. [Online]. Available: https://dial.uclouvain.be/downloader/downloader.php?pid=thesis%3A38837&datastream=PDF_01&cover=cover-mem
14. "What are blockchain games?" Chainlink. Accessed: Jul. 18, 2024. [Online]. Available: https://chain.link/education/blockchain-gaming
15. H. R. Hasan et al., "Non-fungible tokens (NFTs) for digital twins in the industrial metaverse: Overview, use cases, and open challenges," *Comput. Ind. Eng*, vol. 193, 2024, Art. no. 110315, 2024, doi: 10.1016/j.cie.2024.110315.
16. G. Manasa, K. Rajesh, E. L. Goud, G. Shriya, and K. Srijani. "Block chain technology with centralized database for conventional data integrity verification schemes." IJARST. Accessed: Jul. 17, 2024. [Online]. Available: https://ijarst.in/public/uploads/paper/214991716457897.pdf
17. A. Ibrahim, "Guarding the future of gaming: The imperative of cybersecurity," in *Proc. 2nd Int. Conf. Cyber Resilience (ICCR)*, Piscataway, NJ, USA: IEEE Press, 2024, pp. 1–9, doi: 10.1109/ICCR61006.2024.10532843.
18. R. Girasa, Regulation of Cryptocurrencies and Blockchain Technologies: National and International Perspectives (Palgrave Studies in Financial Services Technology). Cham, Switzerland: Springer-Verlag, 2023.
19. Q. Wang, R. Li, Q. Wang, and S. Chen, "Non-fungible token (NFT): Overview, evaluation, opportunities and challenges," Oct. 2021, *arXiv:2105.07447.*
20. R. P. Sarode, Y. Watanobe, and S. Bhalla, "From silos to unity: Seamless cross-platform gaming by leveraging blockchain technology," in *Big Data Analytics in Astronomy, Science, and Engineering* (Lecture Notes in Computer Science), vol. 14516, S. Sachdeva and Y. Watanobe, Eds., Cham, Switzerland: Springer-Verlag, 2024, pp. 213–223.
21. H. Taherdoost and M. Madanchian, "Blockchain-based new business models: A systematic review," *Electronics*, vol. 12, no. 6, pp. 1479, 2023, doi: 10.3390/electronics12061479.
22. "Blockchain gaming market by game type (role playing games, open world games, collectible games), platforms (ETH, BNB chain, polygon), and region (North America, Europe, Asia Pacific, rest of the world) - Global forecast to 2027." Research and Markets. Accessed: Jul. 18, 2024. [Online]. Available: https://www.researchandmarkets.com/reports/5702427/blockchain-gaming-market-by-game-type-role
23. P. Delfabbro, A. Delic, and D. L. King, "Understanding the mechanics and consumer risks associated with play-to-earn (P2E) gaming," *J. Behav. Addict*, vol. 11, no. 3, pp. 716–726, 2022, doi: 10.1556/2006.2022.00066.
24. H. X. Liu and J. P. Holopainen, "Calling for play-oriented research on blockchain video games: An overview study," *Distrib. Ledger Technol. Res. Pract.*, Art. no. 3674154, Jun. 2024, doi: 10.1145/3674154.
25. R. Yang, F. R. Yu, P. Si, Z. Yang, and Y. Zhang, "Integrated blockchain and edge computing systems: A survey, some research issues and challenges," *IEEE Commun. Surv. Tut.*, vol. 21, no. 2, pp. 1508–1532, Secondquarter 2019, doi: 10.1109/COMST.2019.2894727.
26. M. A. Ferrag and L. Shu, "The performance evaluation of blockchain-based security and privacy systems for the Internet of Things: A tutorial," *IEEE Internet Things J.*, vol. 8, no. 24, pp. 17,236–17,260, May 2021, doi: 10.1109/JIOT.2021.3078072.
27. G. Beaumier et al., "Global regulations for a digital economy: Between new and old challenges," *Glob. Policy*, vol. 11, no. 4, pp. 515–522, Sep. 2020, doi: 10.1111/1758-5899.12823.
28. S. Bonnet and F. Teuteberg, "Impact of blockchain and

distributed ledger technology for the management, protection, enforcement and monetization of intellectual property: A systematic literature review," *Inf. Syst. E-Bus. Manage.*, vol. 21, no. 2, pp. 229–275, 2023, doi: 10.1007/s10257-022-00579-y.

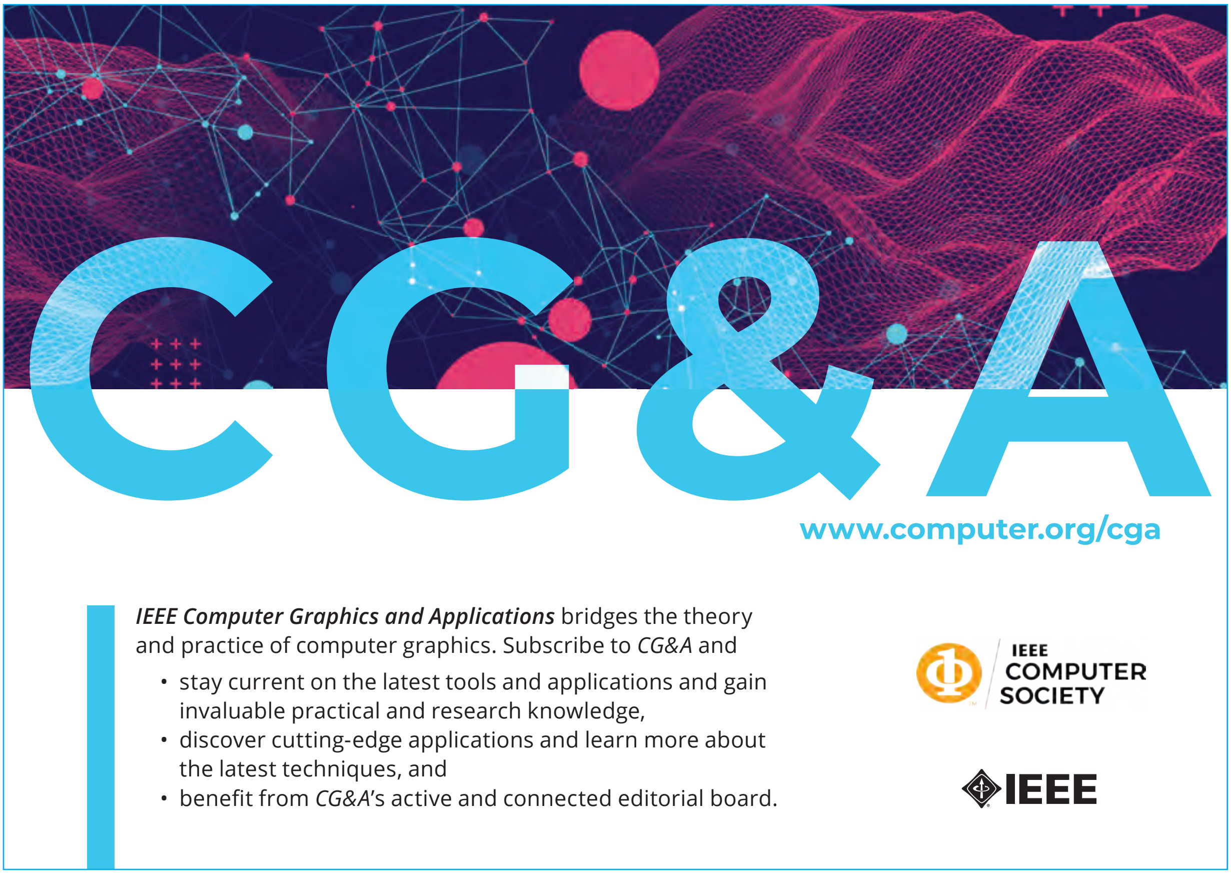
29. A. Averin and A. Samartsev," Review of blockchain in computer games," *AIP Conf. Proc.*, vol. 2910, no. 1, Art. no. 020033.
30. Y. Wang, M. Sheng, and D. A. Ghani, "Virtual reality and augmented reality-based digital pattern design in the context of the blockchain technology framework," *J. Auton. Intell.*, vol. 7, no. 5, pp. 1–11, 2024.
31. A. Alamsyah, G. N. W. Kusuma, and D. P. Ramadhani, "A review on decentralized finance ecosystems," *Future Internet*, vol. 16, no. 3, 2024, Art. no.76, doi: 10.3390/fi16030076.

KEVIN MACWAN is a software development manager for the Fulfillment By Amazon organization. Contact him at kmacwan@usc.edu.

stay on the Cutting Edge of Artificial Intelligence

IEEE Intelligent Systems provides peer-reviewed, cutting-edge articles on the theory and applications of systems that perceive, reason, learn, and act intelligently.

The #1 AI Magazine
www.computer.org/intelligent
IEEE Intelligent Systems



IEEE Computer Graphics and Applications bridges the theory and practice of computer graphics. Subscribe to *CG&A* and

- stay current on the latest tools and applications and gain invaluable practical and research knowledge,
- discover cutting-edge applications and learn more about the latest techniques, and
- benefit from *CG&A*'s active and connected editorial board.

IEEE COMPUTER SOCIETY

IEEE

EDITOR: Nir Kshetri, University of North Carolina at Greensboro, nbkshetr@uncg.edu

DEPARTMENT: COMPUTING'S ECONOMICS

This article originally appeared in Computer vol. 55, no. 8, 2022

Blockchain's Carbon and Environmental Footprints

Nir Kshetri, *University of North Carolina at Greensboro*

Jeffrey Voas, *IEEE Fellow*

We analyze existing discourses surrounding blockchains' energy consumption and look at the actors and actions involved. We also provide an evaluation of various considerations and factors that affect blockchain networks' energy consumption and resulting environmental impacts.

Blockchain networks' energy consumption is a timely topic. According to the Cambridge Bitcoin Electricity Consumption Index, the Bitcoin network consumed 0.61% of world's total electricity production in March 2022. This is more than the total consumption by Ukraine or Norway.[1]

CRYPTOCURRENCIES' PROPONENTS, HOWEVER, HAVE POINTED OUT THAT ELECTRICITY CONSUMED BY BLOCKCHAIN NETWORKS COMPRISES ONLY A SMALL PROPORTION OF THE ELECTRICITY WASTED FROM OTHER SOURCES.

Crypto enthusiasts, policy-making agencies, activists, consumers, and corporations hold divergent perspectives about this. Regulators in China and Kosovo have banned Bitcoin mining. Bitcoin mining's high energy consumption and negative environmental impact have been key reasons. In December 2021, Kosovo imported 40% of its energy. In January 2022, the government decided to ban all cryptomining activities to address the global energy crisis.[2] Environmental activists have campaigned for a complete ban.

Cryptocurrencies' proponents, however, have pointed out that electricity consumed by blockchain networks comprises only a small proportion of the electricity wasted from other sources. Quoting a study of Cambridge Center for Alternative Finance (CCAF), a cointelegraph.com article noted that electricity losses in transmission and distribution in the United States could power the Bitcoin network 2.2 times.[3] Galaxy Digital Mining's study found that the amount of electricity lost in transmission and distribution is approximately 2,205 TWh/year, which is 19.4 times that of the Bitcoin network. Likewise, "always-on" electrical devices in U.S. households consume roughly 1,375 TWh/year, which is 12.1 times that of the Bitcoin network.[4] Hence, it's all relative to where you sit at the table.

ACTORS AND ACTIONS

In some jurisdictions, cryptocurrency has been subjected to increased regulatory scrutiny due to energy supply shortages allegedly created by bitcoin mining activities and perceived adverse environmental impacts. Blackouts have been reported in several cities in countries such as Iran, Kazakhstan, China, and Kosovo. Blackouts have also left thousands of people without power for days.[5]

Regulatory actions have been taken in several jurisdictions. In May 2021, China prohibited the country's financial institutions from engaging in all crypto transactions. This was followed by a ban on cryptocurrency mining in June 2021. In September 2021, the country outlawed cryptocurrencies.[6] One of the main reasons behind the cryptocurrency mining ban was

Digital Object Identifier 10.1109/MC.2022.3176989
Date of current version: 2 August 2022

2469-7087/25 © 2025 IEEE

arguably an increase in illegal coal extraction, which made it difficult to attain China's ambitious environmental goals, and put people's lives in danger. The preliminary investigation of an April 2021 coal mine accident in Xinjiang that trapped 21 people found that the mine was restarted without government permission to meet cryptoserver farms' power demand.[7]

Similarly, in May 2021, the European Central Bank described the "exorbitant carbon footprint" of cryptoassets as "grounds for concern."[8] The European Union (EU) is under pressure from some member states to mitigate negative environmental impacts of blockchain applications. In November 2021, the Swedish government asked the EU to ban "energy-intensive" cryptomining activities.[9]

Likewise, in May 2021, a bill was introduced in the New York State Senate to establish a "moratorium on cryptocurrency mining operations that use proof-of-work (PoW) authentication methods to validate blockchain transactions."[10] In March 2022, the New York State Assembly Environmental Conservation Committee voted to pass the legislation.[11]

Similar concerns have been raised by international developmental organizations.[12] Issuing a warning against El Salvador's Bitcoin Law, which made bitcoin a legal tender effective September 2021, the International Monetary Fund noted that adverse consequences on the environment are among many risks that countries that adopt cryptocurrencies as a national currency or legal tender can face.[13]

Social and environmental activists have played a vocal and visible role in explaining cryptocurrencies' adverse environmental impacts. When cryptocurrency miners started their activities in New York's industrial towns in 2021 using natural gas plants, environmental groups such as Earthjustice and the Sierra Club expressed concerns over the way the cryptomining companies were operating. These groups argued that huge computer farms' operations can increase greenhouse gas emissions and threaten the state's emission-reduction goals, which require more renewable power and reductions in fossil fuel emissions. There are also complaints against using renewable energy. Environmentalists argued that because Bitcoin mining plants can use more energy than most cities, their operations can increase the dependence of others on fossil fuels. And a blogger criticized a permit that allowed a cryptomining firm to draw more than 100 million gallons of water daily from Seneca Lake for cooling purposes. The water would then be returned at a warmer level to a trout stream tributary.[14]

ENVIRONMENTALISTS ARGUED THAT BECAUSE BITCOIN MINING PLANTS CAN USE MORE ENERGY THAN MOST CITIES, THEIR OPERATIONS CAN INCREASE THE DEPENDENCE OF OTHERS ON FOSSIL FUELS.

The environmental organizations that had embraced cryptocurrencies and nonfungible tokens (NFTs) in their fundraising initiatives have been forced to reverse their actions. Nongovernmental environmental organization Greenpeace, which had accepted bitcoin donations since 2014, stopped accepting donations in the cryptocurrency in 2021 due to concerns regarding the amount of energy needed.[15] In February 2022, World Wildlife Fund U.K. tried to raise money with NFTs, specifically what it called *nonfungible animals*, but, facing sharp criticism from traditional conservation supporters, the organization was forced to immediately end sale of the tokens.[16]

Responding to criticisms, defenders of Bitcoin have argued that Bitcoin's environmental impact is significantly lower than that of the financial and banking sectors. One report suggested that the Bitcoin network uses less than half of the energy used by banks' large data centers.[4]

TABLE 1. The key considerations and factors that affect blockchain networks' energy consumption and resulting environmental impacts.

Consideration/factor	Explanation	Example
The ultimate goal of blockchain use	Energy consumption could be more justified for valuable applications of cryptocurrencies or if they are used for good cause.	Although many collectible NFTs have little to no utility, applications such as securing property titles are valuable.
Phase and type of blockchain transactions	Some phases and types of transactions are less energy intensive.	Minting an NFT consumes more energy than transferring ownership.
The source of energy used	Transactions that use renewable energy are more justified.	HIVE claims that it uses only renewable energy to mine Bitcoin and Ether.
Where blockchain applications are carried out	Applications that take advantage of excess energy in some geographic locations can be more justifiable.	Before cryptomining was outlawed, Bitcoin miners in China migrated to locations with abundant hydropower during the rainy season.
Type of blockchain used	Energy consumption can be reduced by using blockchains that rely on PoS consensus model.	OneOf is built on Tezos.

PoS: proof of stake.

Bitcoin proponents have also argued that cryptocurrencies are helping build the future financial system and hence, their benefits outweigh the costs.[15]

CONSIDERATIONS AND FACTORS

A variety of considerations and factors can guide decisions regarding the use of blockchains and potentially minimize the energy use and environmental impacts of blockchain use (see Table 1). Although many collectible NFTs have little to no utility, blockchains can enable valuable applications such as securing property titles. However, whether certain applications of blockchain are good or bad is subjective. Some view blockchain as an opportunity to realize interests and achieve goals that they value highly. A climate activist was quoted as saying that despite high energy consumption and adverse environmental impact, he would support cryptocurrencies as long as they fight the capitalist establishment and take power away from central banks.[9]

DISCLAIMER

The authors are completely responsible for the content in this article. The opinions expressed here are their own.

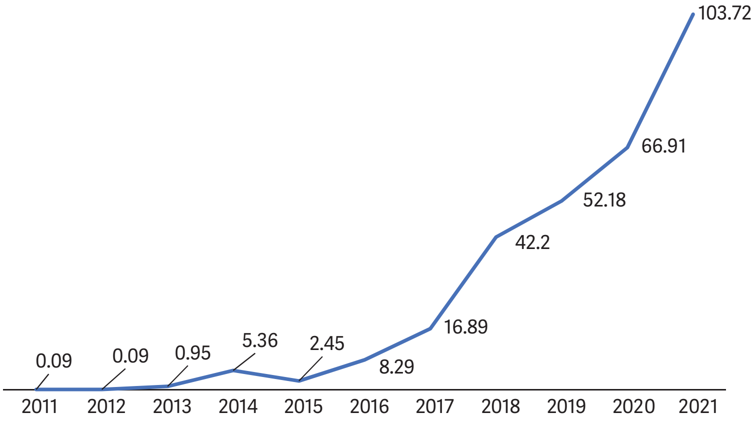
Energy consumption varies across phases and types of transactions. Mining accounts for most of the energy consumption of Bitcoin. For already-mined coins, minimal energy is required to validate transactions.[17] Memo Akten's analysis of 8,000 transactions from the NFT platform SuperRare suggested that an average NFT consumes 340 kWh of energy. According to Akten's calculation, the averages for energy consumption and carbon dioxide (CO_2) emission for different activities associated with NFTs were as follows: minting (creation)—142 kWh, 83 kg CO_2; bids—41 kWh, 24 kg CO_2; cancel bid—12 kWh, 7 kg CO_2; sale—87 kWh, 51 kg CO_2; and transfer of ownership—52 kWh, 30 kg CO_2.[18] Transferring ownership of an already-minted NFT thus creates fewer negative environmental impacts compared to minting a new NFT.

Another consideration is whether the energy used is renewable or not. Bitcoin networks carbon emission level is difficult to estimate with high certainty as miners prefer to hide the details of their operations from competitors. A 2019 report by CoinShares notes that 74% of the world's Bitcoin mining operations "heavily" relied on renewable energy due to the availability of hydropower in mining hubs such as China and Scandinavia.[19] In September 2020, the CCAF estimated renewable energy powered 39% of PoW mining.[20] The proportion further reduced to 25.1% in August 2021 as miners stopped using Chinese hydropower and moved to the United States, where gas supplies much of the power.[21]

TABLE 2. The electricity consumptions of different blockchains.

Blockchain	Annual electricity consumption (TWh)	Blockchain	Annual electricity consumption (TWh)
Bitcoin	204.5	Cardano	0.000598755
Ethereum	112.44	Tezos	0.000113249
Solana	0.01105	Algorand	0.000512671
Polygon	0.00079	Avalanche	0.000489311
Flow	0.00018	—	—

Data source: Bitcoin—estimate by Digiconomist based on annualized data as of 23 March 2022 Ethereum: estimate by Digiconomist based on annualized data as of 23 March 2022,[27] Solana,[28] Polygon,[29] Flow,[30] and Cardano, Algorand, Tezos, and Avalanche.[31]



Data Source[29]

FIGURE 1. The bitcoin network's energy consumption (TWh).[1]

Some bitcoin miners are positioning themselves as environmentally responsible. Canada-based HIVE Blockchain Technologies, which was listed on Nasdaq in 2021, claimed that it uses only renewable energy to mine Bitcoin and Ether.[22] Some critics, however, have questioned the justifiability of using energy, whether renewable or nonrenewable, to power energy-intensive applications such as Bitcoin mining. They suggest that the argument that Bitcoin's high energy consumption and environmental burden can be compensated for by plugging into renewable sources is convenient but possibly false. The renewable resources used to power blockchains could be deployed to more essential needs.[23]

Another way to reduce the environmental impact is to take advantage of arbitrage geographic opportunities, that is, moving activities across borders to utilize excess energy production. This is possible because blockchains' energy consumption differs from most other industries; whereas energy used for other activities must be produced close to its end users, bitcoin can be mined anywhere in the world. In this way, miners can utilize power sources that cannot be used by other applications.[20] Before cryptomining was outlawed in China, bitcoin miners used to migrate to the mountainous provinces with abundant hydropower resources during the rainy season. In these provinces, they took advantage of the excess electricity for several months each year.[24]

Finally, energy consumption and environmental impacts vary across the types of blockchain networks. The blockchains that rely on PoW consensus mechanisms consume more energy (Table 2). Moreover, the energy consumption of these networks is growing rapidly (Figure 1). By using blockchains based on the proof-of-stake (PoS) consensus model, in which only a small group of nodes can validate transactions, energy consumption can be reduced. Some platforms advertise lower energy consumption as a selling proposition. The NFT platform designed for the music industry is built on Tezos,[25] and OneOf promotes itself as a sustainable company.

Cryptocurrencies' high energy consumption is a basis for regulatory scrutiny. More energy-efficient blockchains exist that run on PoS algorithms, but their use has been limited because they lack the characteristics of completely decentralized blockchains.

Whether high energy consumption is viewed as justifiable or not depends on whether we value the functions and services blockchain provides. The question of whether millions of dollars should be spent on an NFT that consumes 340 kWh of electricity is a question of values. The individuals that consider cryptocurrencies to be a tool to build future financial systems and fight capitalism may view this energy consumption as justifiable. On the other hand, those that view cryptocurrencies as a "fraud" or "Ponzi scheme" may consider this energy consumption a waste.

Measures can be taken to mitigate the high energy consumption and adverse environmental impacts. Blockchain applications such as Bitcoin mining and minting NFTs can be performed throughout the world. The environmental impacts can thus be reduced if these activities are performed in locations with excess energy. Likewise, blockchain activities that employ renewable energy may be more justified due to their carbon-neutral nature.

REFERENCES

1. "Bitcoin network power demand," Cambridge Centre for Alternative Finance, Cambridge, U.K. Accessed: May 15, 2022. [Online]. Available: https://ccaf.io/cbeci/index
2. "Kosovo bans cryptocurrency mining after blackouts." BBC. https://www.bbc.com/news/world-europe-59879760 (Accessed: May 15, 2022).
3. M. Van Niekerk. "Enterprise blockchain to play a pivotal role in creating a sustainable future." Cointelegraph. https://cointelegraph.com/news/enterprise-blockchain-to-play-a-pivotal-role-in-creating-a-sustainable-future (Accessed: May 15, 2022).
4. R. Rybarczyk, D. Armstrong, and A. Fabiano. "On bitcoin's energy consumption: A quantitative approach to a subjective question." Galaxy Digital. https://www.lopp.net/pdf/On_Bitcoin_Energy_Consumption.pdf (Accessed: May 15, 2022).
5. G. Bandera. "Is cryptocurrency bad for the environment?" Fairplanet. https://www.fairplanet.org/story/is-cryptocurrency-bad-for-the-environment/ (Accessed: May 15, 2022).
6. M. Quiroz-Gutierrez. "Crypto is fully banned in China and 8 other countries." Fortune. https://fortune.com/2022/01/04/crypto-banned-china-other-countries/#:~:text=When%20it%20banned%20crypto%20last%20year%2C%20China%20did%20so%20in,outlawed%20cryptocurrencies%20outright%20in%20September (Accessed: May 15, 2022).
7. "China's latest crackdown on crypto caused by climate concerns." Aljazeera. https://www.aljazeera.com/economy/2021/5/26/bbchinas-latest-crackdown-on-crypto-caused-by-surge-in-coal-mini (Accessed: May 15, 2022).
8. "Financial stability review," European Central Bank, Frankfurt am Main, Germany, May 2021. [Online]. Available: https://www.ecb.europa.eu/pub/financial-stability/fsr/html/ecb.fsr202105~757f727fe4.en.html
9. S. Mellor. "How crypto-owning climate activists balance saving the planet with supporting energy-hungry Bitcoin mines." Fortune. https://fortune.com/2021/11/15/cop26-cryptocurrency-owning-climate-activists-bitcoin-blockchain-energy-use/ (Accessed: May 15, 2022).
10. "Senate bill S6486D 2021-2022 legislative session." NY State Senate. https://www.nysenate.gov/legislation/bills/2021/s6486 (Accessed: May 15, 2022).
11. "NY crypto mining moratorium passes assembly environmental conservation committee: Environmental community looks to Sen. Parker for companion legislation." Food & Water Watch. https://www.foodandwaterwatch.org/2022/03/22/ny-crypto-mining-moratorium-passes-assembly-environmental-conservation-committee/ (Accessed: May 15, 2022).
12. N. Kshetri, "El Salvador's bitcoin gamble," *Computer*, vol. 55, no. 6, pp. 85–89, 2022.
13. T. Wright. "IMF issues veiled warning against El Salvador's Bitcoin Law." Cointelegraph. https://cointelegraph.com/news/imf-issues-veiled-warning-against-el-salvador-s-bitcoin-law (Accessed: May 15, 2022).
14. C. Kilgannon, "Quotation of the day: The climate cost of a Bitcoin boom," *New York Times*. [Online]. Available: https://www.nytimes.com/2021/12/07/todayspaper/quotation-of-the-day-the-climate-cost-of-a-bitcoin-boom.html (Accessed: May 15, 2022).
15. K. Martin and B. Nauman, "Bitcoin's growing energy problem: 'It's a dirty currency'," *Financial Times*. [Online]. Available: https://www.ft.com/content/1aecb2db-8f61-427c-a413-3b929291c8ac (Accessed: May 15, 2022).
16. "WWF-UK ends sale of NFTs after backlash, angering the crypto community," *Climate Home News*. [Online]. Available: https://www.climatechangenews.com/2022/02/09/wwf-uk-ends-sale-nfts-backlash-angering-crypto-community/ (Accessed: May 15, 2022).
17. N. Carter, How much energy does bitcoin actually consume? *Harvard Bus*. Rev. [Online]. Available: https://hbr.org/2021/05/how-much-energy-does-bitcoin-actually-consume (Accessed: May 15, 2022).
18. A. Storey. "How much energy does it take to make an

NFT?" Poster Grind. https://postergrind.com/how-much-energy-does-it-take-to-make-an-nft/ (Accessed: May 15, 2022).

19. J. Redman. "74% of the world's bitcoin mining operations driven by renewable energy says report." Bitcoin.com. https://news.bitcoin.com/74-of-the-worlds-bitcoin-mining-operations-driven-by-renewable-energy-says-report/ (Accessed: May 15, 2022).
20. A. Blandin *et al.*, "3rd global cryptoasset benchmarking study," Cambridge Centre for Alternative Finance, Cambridge, U.K., Sep. 2020. [Online]. Available: https://www.jbs.cam.ac.uk/wp-content/uploads/2021/01/2021-ccaf-3rd-global-cryptoasset-benchmarking-study.pdf
21. "Bitcoin less green since China ban, research suggests." BBC. https://www.bbc.com/news/technology-60521975 (Accessed: May 15, 2022).
22. "Bitcoin mining uses a higher mix of sustainable energy than any major country or industry," *Forbes*. [Online]. Available: https://www.forbes.com/sites/greatspeculations/2021/07/06/bitcoin-mining-uses-a-higher-mix-of-sustainable-energy-than-any-major-country-or-industry/?sh=59c207a84cc9 (Accessed: May 15, 2022).
23. J. Baguley, "Letter: Plugging blockchain into green energy is no solution," *Financial Times*. [Online]. Available: https://www.ft.com/content/f3f259f3-952f-46fc-b6f4-840add3c718b (Accessed: May 15, 2022).
24. J. Coroneo-Seaman. "'Great mining migration': Power-hungry Bitcoin leaves China." China Dialogue. https://chinadialogue.net/en/enegy/great-mining-migration-power-hungry-bitcoin-leaves-china/ (Accessed: May 15, 2022).
25. N. Rubio-Licht, "OneOf plans to make affordable NFTs for musicians," *Los Angeles Bus. J.* [Online]. Available: https://labusinessjournal.com/news/2021/jun/14/oneof-make-affordable-nfts-musicians/ (Accessed: May 15, 2022).
26. "Bitcoin energy consumption index." Digiconomist. https://digiconomist.net/bitcoin-energy-consumption/ (Accessed: May 15, 2022).
27. "Ethereum energy consumption index." Digiconomist. https://digiconomist.net/ethereum-energy-consumption/ (Accessed: May 15, 2022).
28. "Solana's energy use report: November 2021." Solana. https://solana.com/news/solana-energy-usage-report-november-2021 (Accessed: May 15, 2022).
29. "Polygon: The eco-friendly blockchain scaling Ethereum." Polygon. https://blog.polygon.technology/polygon-the-eco-friendly-blockchain-scaling-ethereum-bbdd52201ad/ (Accessed: May 15, 2022).
30. "New findings from Deloitte Canada reveal minting an NFT on Flow takes less energy than a Google search or Instagram post." Flow. https://www.onflow.org/post/flow-blockchain-sustainability-energy-deloitte-report-nft (Accessed: May 15, 2022).
31. K. De Ceunynck. "Evaluating Tezos: Energy consumption and carbon footprint. Cryptomode." https://cryptomode.com/evaluating-tezos-energy-consumption-and-carbon-footprint/#:~:text=Electricity%20consumption%20per%20transaction&text=Tezos%20consumes%2041.45%20Wh%2Ftx%20per%20network (Accessed: May 15, 2022).

NIR KSHETRI is a professor at the Bryan School of Business and Economics at the University of North Carolina at Greensboro, Greensboro, North Carolina, 27412, USA, and the "Computing's Economics" column editor for *Computer*. Contact him at nbkshetr@uncg.edu.

JEFFREY VOAS, Gaithersburg, Maryland, USA, is the editor in chief of *Computer*. He is a Fellow of IEEE. Contact him at j.voas@ieee.org.

IEEE MultiMedia serves the community of scholars, developers, practitioners, and students who are interested in multiple media types and work in fields such as image and video processing, audio analysis, text retrieval, and data fusion.

Read It Today!

www.computer.org/multimedia

IEEE COMPUTER SOCIETY IEEE

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

Get Published in the *IEEE Open Journal of the Computer Society*

Get more citations by publishing with the *IEEE Open Journal of the Computer Society*

Your research on computing and informational technology will benefit from 5 million unique monthly users of the *IEEE Xplore*® Digital Library. Plus, this journal is fully open and compliant with funder mandates, including Plan S.

Submit your paper today!
Visit www.computer.org/oj to learn more.

IEEE COMPUTER SOCIETY | IEEE

IEEE COMPUTER SOCIETY

Publications Seek 2026 Editors in Chief

Application Deadline: 1 March 2025

IEEE Computer Society seeks applicants for editor in chief for the following publications:

- *Computer* magazine
- *IEEE Computer Graphics and Applications*
- *IEEE Security & Privacy*
- *IEEE Transactions on Emerging Technologies in Computing*
- *IEEE Transactions on Mobile Computing*
- *IEEE Transactions on Services Computing*
- *IEEE Transactions on Software Engineering*
- *IEEE Transactions on Visualization and Computer Graphics*

Our publications are the cornerstone of professional activities for our members and the community we serve. We seek candidates who are IEEE members in good standing, have strong familiarity with our publications, and possess an excellent understanding of the field as it relates to academic, industry, and governmental areas. Applicants must have successful experience developing a diverse team of individuals to serve key editorial board roles. Demonstrated managerial skills are also required to ensure content and issue development, and timely processing of submissions. Terms begin 1 January 2026.

For complete information on how to apply, please go to
www.computer.org/press-room/seeking-2026-editors-in-chief

Apply Today!

AI's 10 to Watch
Deep Anomaly Analytics
Computer Graphics
Multisensory Experiences in eXtended Reality
A Blockchain-Based Copyright Protection Scheme With Proactive Defense
Simulation Driven AI: From Artificial to Actual and Vice Versa
SECURITY & PRIVACY
Inclusive Privacy and Security
COMPETITIONS
Computer
Trustworthy AI—Part II
A Flexible Heuristic to Schedule Distributed Analytic Applications in Compute Clusters
SPOTLIGHT ON TRANSACTIONS
Contributions of Quantum Factoring on Quantum Research

IEEE COMPUTER SOCIETY

IEEE

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.

FEBRUARY

9 February

- BigComp (IEEE Int'l Conf. on Big Data and Smart Computing), Kota Kinabalu, Malaysia

17 February

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

26 February

- VISIGRAPP (Int'l Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications), Porto, Portugal

28 February

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

MARCH

1 March

- HPCA (IEEE Int'l Symposium on High Performance Computer Architecture), Las Vegas, USA

4 March

- SANER (IEEE Int'l Conf. on Software Analysis, Evolution and Reengineering), Montreal, Canada

8 March

- VR (IEEE Conf. Virtual Reality and 3D User Interfaces), Saint Malo, France

17 March

- PerCom (IEEE Int'l Conf. on Pervasive Computing and Communications), Washington, DC, USA

31 March

- DATE (Design, Automation & Test in Europe Conf.), Lyon, France
- ICSA (IEEE Int'l Conf. on Software Architecture), Odense, Denmark
- ICST (IEEE Conf. on Software Testing, Verification and Validation), Naples, Italy

APRIL

9 April

- SaTML (IEEE Conf. on Secure and Trustworthy Machine Learning), Copenhagen, Denmark

15 April

- CSASE (Int'l Conf. on Computer Science and Software Eng.), Duhok, Iraq

16 April

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

22 April

- PacificVis (IEEE Pacific Visualization Conf.), Taipei City, Taiwan

26 April

- ICSE (IEEE/ACM Int'l Conf. on Software Eng.), Ottawa, Canada

28 April

- VTS (IEEE VLSI Test Symposium), Tempe, USA

MAY

4 May

- ARITH (IEEE Symposium on Computer Arithmetic), El Paso, USA
- FCCM (IEEE Annual Int'l Symposium on Field-Programmable Custom Computing Machines), Fayetteville, USA
- MOST (IEEE Int'l Conf. on Mobility, Operations, Services and Technologies), Newark, USA

5 May

- CAI (IEEE Conf. on Artificial Intelligence), Santa Clara, USA
- HOST (IEEE Int'l Symposium on Hardware Oriented Security and Trust), San Jose, USA

6 May

- RTAS (IEEE Real-Time and Embedded Technology and Applications Symposium), Irvine, USA

11 May

- ASYNC (IEEE Int'l Symposium on Asynchronous Circuits and Systems), Portland, USA
- ISPASS (IEEE Int'l Symposium on Performance Analysis of Systems and Software), Ghent, Belgium

12 May

- SP (IEEE Symposium on Security and Privacy), San Francisco, USA

19 May

- CCGrid (IEEE Int'l Symposium on Cluster, Cloud and Internet Computing), Tromsø, Norway
- ICDE (IEEE Int'l Conf. on Data Eng.), Hong Kong
- ICFEC (IEEE Int'l Conf. on Fog and Edge Computing) Tromsø, Norway

26 May

- FG (IEEE Int'l Conf. on Automatic Face and Gesture Recognition), Tampa/Clearwater, USA

JUNE

2 June

- MDM (IEEE Int'l Conf. on Mobile Data Management), Irvine, USA

3 June

- IPDPS (IEEE Int'l Parallel and Distributed Processing Symposium), Milano, Italy

5 June

- ISMVL (IEEE Int'l Symposium on Multiple-Valued Logic), Montreal, Canada

11 June

- CVPR (IEEE/CVF Conf. on Computer Vision and Pattern Recognition), Nashville, USA

16 June

- CSF (IEEE Computer Security Foundations Symposium), Santa Cruz, USA

18 June

- CBMS (IEEE Int'l Symposium on Computer-Based Medical Systems), Madrid, Spain
- ICHI (IEEE Int'l Conf. on Healthcare Informatics), Rende, Italy

21 June

- ISCA (ACM/IEEE Annual Int'l Symposium on Computer Architecture), Tokyo, Japan

23 June

- DSN (Annual IEEE/IFIP Int'l Conf. on Dependable Systems and Networks), Naples, Italy
- SVCC (Silicon Valley Cybersecurity Conf.), San Francisco, USA

26 June

- IEEE Cloud Summit, Washington, DC, USA

30 June

- EuroS&P (IEEE European Symposium on Security and Privacy), Venice, Italy
- ICME (IEEE Int'l Conf. on Multimedia and Expo), Nantes, France

JULY

6 July

- ISVLSI (IEEE Computer Society Annual Symposium on VLSI), Kalamata, Greece

7 July

- IOLTS (IEEE Int'l Symposium on On-Line Testing and Robust System Design), Ischia, Italy
- SERVICES (IEEE World Congress on Services), Helsinki, Finland

8 July

- COMPSAC (IEEE Annual Computers, Software, and Applications Conf.), Toronto, Canada

14 July

- ICALT (IEEE Int'l Conf. on Advanced Learning Technologies), Changhua, Taiwan

20 July

- ICDCS (IEEE Int'l Conf. on Distributed Computing Systems), Glasgow, United Kingdom

21 July

- ICCP (IEEE Int'l Conf. on Computational Photography), Toronto, Canada

IEEE COMPUTER SOCIETY CAREER CENTER

Career Accelerating Opportunities

Explore new options—upload your resume today

careers.computer.org

Changes in the marketplace shift demands for vital skills and talent. The IEEE Computer Society Career Center is a valuable resource tool to keep job seekers up to date on the dynamic career opportunities offered by employers.

Take advantage of these special resources for job seekers:

- JOB ALERTS
- TEMPLATES
- WEBINARS
- CAREER ADVICE
- RESUMES VIEWED BY TOP EMPLOYERS

No matter what your career level, the IEEE Computer Society Career Center keeps you connected to workplace trends and exciting career prospects.

IEEE COMPUTER SOCIETY

IEEE