Please cite this article in press as: Lee et al., A view of the sustainable computing landscape, Patterns (2025), https://doi.org/10.1016/j.patter.2025.101296

Patterns

CellPress
OPEN ACCESS

Perspective

A view of the sustainable computing landscape

Benjamin C. Lee,[1,*] David Brooks,[2,*] Arthur van Benthem,[1] Mariam Elgamal,[2] Udit Gupta,[3] Gage Hills,[2] Vincent Liu,[1] Linh Thi Xuan Phan,[1] Benjamin Pierce,[1] Christopher Stewart,[4] Emma Strubell,[5] Gu-Yeon Wei,[2] Adam Wierman,[6] Yuan Yao,[7] and Minlan Yu[2]

[1]University of Pennsylvania, Philadelphia, PA, USA
[2]Harvard University, Cambridge, MA, USA
[3]Cornell University, Ithaca, NY, USA
[4]Ohio State University, Columbus, OH, USA
[5]Carnegie Mellon University, Pittsburgh, PA, USA
[6]California Institute of Technology, Pasadena, CA, USA
[7]Yale University, New Haven, CT, USA
*Correspondence: leebcc@seas.upenn.edu (B.C.L.), dbrooks@g.harvard.edu (D.B.)
https://doi.org/10.1016/j.patter.2025.101296

THE BIGGER PICTURE The digital world is expanding at an unprecedented pace, driven by the explosive growth of artificial intelligence, data center computing, and networked devices. However, these computing technologies come with an environmental cost that is also growing rapidly but is still poorly understood. These costs arise from the energy-intensive data centers that power artificial intelligence as well as from the manufacture of semiconductors that store and compute on massive datasets. Current trends are unsustainable as artificial intelligence transforms the way we live and work and, consequently, as the demand for computing accelerates.

This article presents an agenda for making computation more sustainable by rethinking how we design, build, and operate digital systems. The agenda is interdisciplinary and spans hardware design, software optimization, energy systems, and economic policy. It seeks to mitigate both embodied carbon, the emissions associated with manufacturing hardware like chips and servers, and operational carbon, the emissions associated with the electricity used to power this hardware. Reducing both types of emissions will require modular hardware organizations that allow greater reuse, energy-efficient data center design and management, and intelligent use of renewable or carbon-free energy. The authors encourage collaboration across disciplines—from computer science and engineering to economics and environmental science—to ensure that technical solutions align with societal goals.

SUMMARY

This article presents a holistic research agenda to address the significant environmental impact of information and communication technology (ICT), which accounts for 2.1%–3.9% of global greenhouse gas emissions. It proposes several research thrusts to achieve sustainable computing: accurate carbon accounting models, life cycle design strategies for hardware, efficient use of renewable energy, and integrated design and management strategies for next-generation hardware and software systems. If successful, the research would flatten and reverse growth trajectories for computing power and carbon, especially for rapidly growing applications like artificial intelligence. The research takes a holistic approach because strategies that reduce operational carbon may increase embodied carbon, and vice versa. Achieving these goals will require interdisciplinary collaboration between computer scientists, electrical engineers, environmental scientists, and economists.

INTRODUCTION

Information and communication technology (ICT) accounts for a surprisingly large share of global greenhouse gas (GHG) emissions—estimates range from 2.1% to 3.9%. To tackle this challenge, the International Telecommunication Union aims for a 45% reduction in ICT emissions by 2030,[1] aligning with the Paris Agreement's goal to limit warming to 1.5°C above pre-industrial levels. Meeting the growing demands for computing while achieving these goals will be difficult and costly, requiring rigorous methods that balance sustainability benefits against implementation costs. To succeed, computer scientists, electrical engineers, environmental scientists, and economists must develop an ecosystem for sustainable computing with

Please cite this article in press as: Lee et al., A view of the sustainable computing landscape, Patterns (2025), https://doi.org/10.1016/j.patter.2025.101296

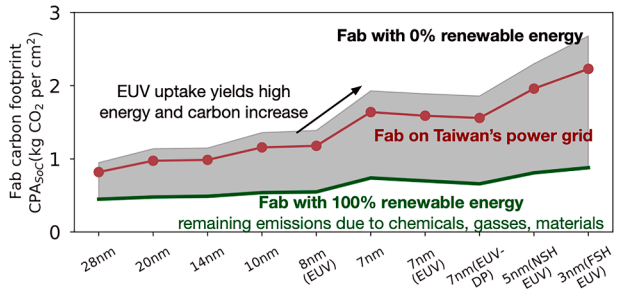

Figure 1. Embodied carbon for semiconductor fabrication
Data from industry reports, device characterization.[8]

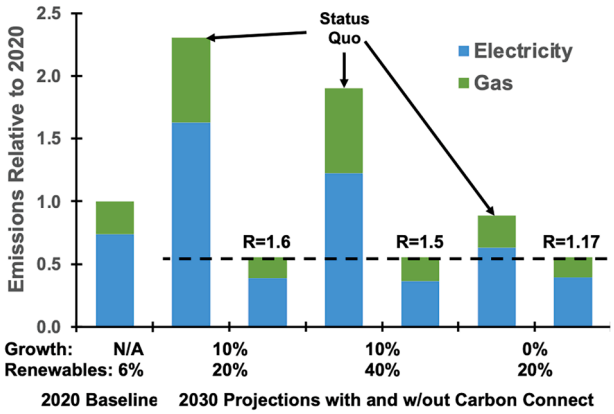

Figure 2. Embodied carbon scenarios that vary fab electricity growth, renewable energy use, characterization, and 3Rs of circular economy

transformative solutions to computing's carbon problem. This responds to the call for action from Knowles et al.[2]: computing must end the "digital exceptionalism" that overlooks its carbon footprint due to its contributions to societal productivity and efficiency.

We envision several interlocking research thrusts to address these sustainability challenges for next-generation computer systems. These thrusts take a coordinated approach to hardware and software, designing new processors, servers, and data centers as well as optimizing their deployment for emerging artificial intelligence (AI) applications. These thrusts also take a holistic view of computing's carbon footprint, reducing embodied carbon from hardware manufacturing via life cycle design strategies and reducing operational carbon from the judicious, timely use of carbon-efficient electricity. We anticipate interesting trade-offs between embodied and operational carbon, as a solution might reduce one type of carbon at the expense of the other. Finally, these solutions must account for the broader economic and policy context to align private initiatives with societal goals.

This article briefly surveys the challenges and opportunities in sustainable computing. It reflects the research priorities of the authors, but the holistic perspective may inspire researchers from diverse intellectual communities—computer science, electrical engineering, industrial ecology, economics, and law—to engage with these questions. We recognize that some of these research questions are becoming qualitatively more challenging due to interest in AI and investment in hyperscale data centers. We also recognize that some of these questions, such as life cycle analysis for hardware, are benefiting from industry attention. This article seeks to place these recent developments in context and encourage greater coordination between these individual research contributions.

EMBODIED CARBON

Embodied carbon describes emissions associated with computing's demands on hardware manufacturing and supply chains; the GHG Protocol designates these as scope 3 emissions.[3–5] These costs are significant for high-performance computing due to unprecedented data center construction and massive capital investments in graphics processing units and other hardware components for AI. They are also significant for embedded and mobile devices due to high replacement rates and relatively low utilization. Nearly 75% of Apple's emissions are due to manufacturing.[4] Billions of devices are expected to come online by 2027, and their embodied carbon may approach one gigaton of CO_2 per year, exceeding commercial aviation's footprint.[5]

Semiconductor fabrication's contributions to global warming are attributed to electricity and gasses used in manufacturing. Electricity use is particularly significant for advanced technology nodes that require extreme ultraviolet lithography (Figure 1). Carbon-free electricity is a meager 6% of the total in Taiwan and South Korea, where most chips are produced, but the Taiwan Semiconductor Manufacturing Company (TSMC) and Korea may increase their use of carbon-free energy to 40% and 20% of their respective totals by 2030.[6,7]

Figure 2 presents several scenarios for embodied carbon. Even under optimistic assumptions where fab demand is unchanged (0%) and the renewable energy supply increases by 20%, the industry will miss its goal of reducing emissions by 45%, as indicated by the dashed line in the figure. This outcome is partially explained by gases, which account for 25% of total emissions and are unaffected by the use of renewable energy. Thus, reducing embodied carbon by 45% requires more aggressive, innovative measures.

Researchers will need to explore several mitigation strategies that arise from the Rs of the circular economy—reduce, reuse, and recycle. Our analysis specifies an "R factor" that estimates the extent to which these Rs are needed to reduce embodied carbon by 45%. For example, R = 1.5 estimates the combined effect of reducing hardware procurement by 33%, reusing hardware 1.5× longer, and recycling 1.5× more hardware relative to 2020 levels. While different combinations are possible, increasing each of the three Rs is essential for the 45% reduction target.

Reduce

Computer architects should precisely manufacture, provision, and allocate the hardware required for software needs. We need hardware functions that can be designed and implemented separately as small chiplets and then connected with fast networks.[9] Chiplets are more carbon efficient, as fabs precisely manufacture the required circuits and no more, reducing the silicon area and improving manufacturing yields, which in turn reduces waste and carbon. Moreover, fabs could separate the

Please cite this article in press as: Lee et al., A view of the sustainable computing landscape, Patterns (2025), https://doi.org/10.1016/j.patter.2025.101296

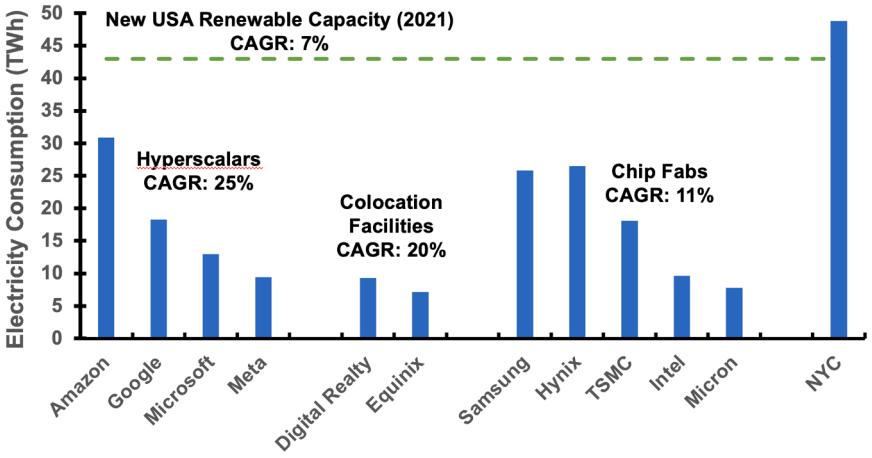

Figure 3. Electricity usage (2021) for data center and fabrication facilities
Compound annual growth rate from 2015 to 2021. Corporate sustainability reports and EIA.[4]

manufacture of disparate capabilities—compute, memory, and sensors—and use dedicated process flows for each, reducing the number of process steps and associated carbon.

We also need data-center-scale disaggregation, which organizes hardware into collections of network-attached components. Compute nodes would offer many central processing units (CPUs) but little dynamic random access memory (DRAM), whereas memory nodes would offer the reverse. Disaggregation allows servers to independently scale a specific hardware type. "Lego-block" systems with custom core and memory configurations would better balance the system and improve carbon efficiency. Today's servers provision many DRAMs for capacity but must also inefficiently provision a corresponding number of memory channels and processor sockets even when workloads under-utilize these channels and processors.[10]

Reuse

Data center operators might replace hardware components based on individual technology advances or failure rates rather than on the fastest evolving or least reliable component, thereby extending the hardware's average tenure. For example, graphics processing units (GPUs) might refresh at a rate dictated by growing demands for AI workloads, whereas CPUs might refresh at a different rate, tracking demand for general computation. Today, the typical server lifetime is 3–6 years, after which the entire rack is replaced with new hardware. Networking equipment lifetimes are longer, 5 years for switches/routers and 10 years for the fiber cable plant, but periodic and wholesale replacement is still common.

Recycle

Hardware will require better instrumentation and health models to facilitate an efficient secondary market that disassembles systems into components and sells them for a second life. For instance, heavily used processors from data centers will have very different resale values than lightly used ones from enterprises. Hardware "odometers" could be implemented with immutable, tamper-resistant registers that count operations. For memories, registers might count errors and faults as well as reads and writes. Measures of physical conditions such as power variations, thermal stresses, and humidity will be helpful. These data must be curated by manufacturers, sellers, or third parties so that consumers can intelligently assign value to pre-owned hardware. We draw inspiration from the role that odometers, vehicle history reports, and certified pre-owned designations play in the secondary vehicle market.

OPERATIONAL CARBON

Operational carbon describes emissions associated with computing's electricity use; the GHG protocol designates these as scope 2 emissions. These costs exhibit explosive growth, driven by AI and its myriad applications. Annual ICT energy demand is projected to exceed 100 exajoules, nearly 15% of the world's energy production.[11] Electricity use at Google, Meta, and Microsoft grew at a compound annual growth rate (CAGR) of 25% per year from 2015 to 2021, nearly quadrupling. In contrast, US renewable energy investments grew by only 7% per year (Figure 3). In 2021, hyperscale data centers consumed 19 TWh more than in 2020, nearly half of the 44 TWh of new renewable capacity.

Our analysis highlights the essential role of renewable energy in computing (Figure 4). If renewable energy capacity grows at 10% per year, as forecasted by the US Energy Information Administration (EIA), and computing's energy demand remains at 2020 levels, carbon emissions would fall by 36%. However, Figure 5 indicates computing's energy demand may increase by 10%–25% per year based on forecasts by industry groups[11] and various consultancies.[12–16]

Carbon-free energy growth would struggle to keep pace. Meeting these demands yet reducing carbon by 45% requires computing to adopt renewable energy at 1.7–1.9× faster than

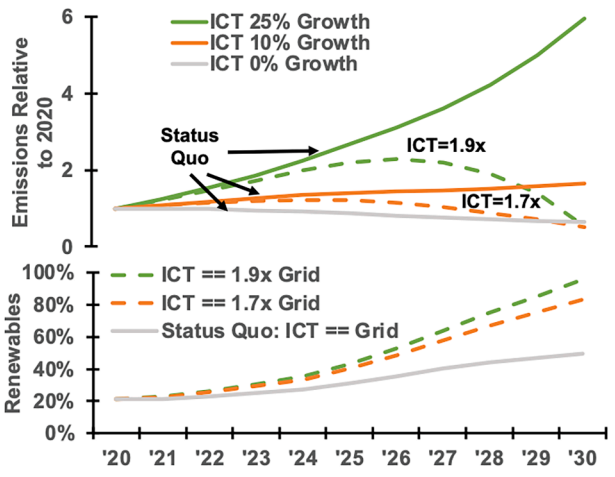

Figure 4. Operational carbon reduction (45% by 2030) achieved via 1.7× higher uptake in ICT renewable electricity compared to the grid

Please cite this article in press as: Lee et al., A view of the sustainable computing landscape, Patterns (2025), https://doi.org/10.1016/j.patter.2025.101296

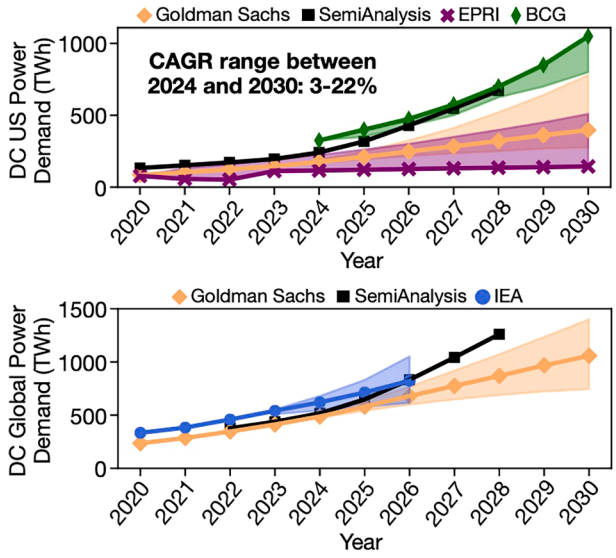

Figure 5. Electricity usage forecasts for data center power in the US and globally
Variance in CAGR estimates is significant.

the US average. Nuclear, whether refurbishing existing plants or building small modular reactors, could be a carbon-free alternative. But there is great uncertainty in this pathway, as the nuclear industry must show that it can build capacity on schedule and within budget in the US.

Demand response

As renewable energy proliferates, sustainable data centers must delay or boost computation based on the availability of carbon-free energy.[17–19] Such demand response (DR) requires grid and data center coordination. One interface would use real-time prices to incentivize data centers to modulate energy use, but this departs from today's contracts that charge based on the amount of power provisioned rather than used. An alternative interface would simply communicate carbon intensity, assuming data centers would modulate demand without compensation.

DR will require hardware and software to trade off performance and power. Ideally, DR frameworks will incentivize participation and guarantee service. Game theory could model system dynamics when users selfishly pursue performance goals. Real-time scheduling and robust machine learning could ensure decisions satisfy diverse obligations. Ultimately, DRs require rethinking conventional wisdom in which data centers constantly compute at peak power to amortize facility and power costs.[20]

Power modulation

Each user must define and implement multiple operating modes that modulate power when required. Hardware mechanisms will rely on energy proportionality, the idea that power should rise and fall with workload. Energy-proportional hardware is difficult to design because most components have a significant fixed power cost dissipated even at near-zero loads. Decades of research have improved CPUs, but today's data centers deploy large memory systems and graphics processing units that will need to be designed for energy proportionality.

Software mechanisms will rely on approximate, degraded computing. Online applications implement contingency plans for site events, ensuring varying degrees of service that depend on system availability and downtime. We will explore real-time system design and anytime algorithms to provide a smoother spectrum of trade-offs between quality and power than permitted in today's systems. Strategies for computational sprinting might allow workloads to dynamically consume additional resources as power budgets permit.[21]

Intelligent decisions

A cognitive stack could organize power management into a low-level reactive layer and a high-level deliberative layer. An agent monitors software performance and hardware utilization, optimizing power use to achieve performance goals while accounting for data center conditions and competition from other agents. The reactive policy would adjust a processor's power use based on program phases, while the deliberative policy would ensure that adjustments align with other processors' policies and data center goals in sustainability, safety, and stability.

The cognitive stack could use multi-agent game theory and reinforcement learning for dynamic decision-making.[22] Dynamism is crucial because computation varies over time, and allocation decisions in the present should account for the past and anticipate the future. For example, in a repeated game, agents spend tokens for power and learn policies for spending, requesting power, and using hardware. When carbon-free energy is scarce, data centers could offer tokens to jobs that defer their computation or require more tokens from those that do not. How should agents spend tokens to maximize long-term performance when allocations in one time period affect those in an uncertain future? How should data centers price power to achieve sustainability goals?

DRIVING APPLICATIONS

AI will drive increasingly rapid growth in computing. Training requires hundreds of thousands of processors that collaboratively consume massive datasets and compute for weeks or months to compute parameter values for a model. Inference requires a rapidly growing number of processors that invoke trained models and respond to user or application prompts, often with ambitious goals for accuracy, response time (i.e., latency), and response rate (i.e., throughput). Efficient AI requires software solutions, such as specialized models that compute equally accurate answers with fewer calculations,[23] and hardware solutions, such as application-specific integrated circuits, that reduce the cost of each calculation.

Advances in AI are enabled by scaling deep models and their training data,[24] which impacts sustainability.[25,26] Benchmarking AI's carbon footprint would help researchers identify the most pressing challenges.[27] An integrated hardware-software perspective will be particularly helpful as researchers explore the net impact of custom hardware,[28] which reduces operational carbon through energy efficiency but increases embodied carbon through semiconductor manufacturing.

Sustainable AI hinges on its responsiveness to the varying availability of data, hardware, and electricity. We will need to design, train, and deploy AI models that offer performance and

Please cite this article in press as: Lee et al., A view of the sustainable computing landscape, Patterns (2025), https://doi.org/10.1016/j.patter.2025.101296

efficiency on a broad spectrum of hardware platforms. Such models would not only ensure backward compatibility for and equity of access to AI features, but they might also slow the rate of hardware refreshes. How can we develop models and platforms that remain relevant over longer periods and better amortize the carbon costs of model training?

There is a complementary need for programmable, reconfigurable hardware that supports a broad spectrum of AI workloads. Such processors would allocate precisely the hardware required for data processing, training, or inference, consuming energy in proportion to utilization. Instead of designing static AI accelerators, how can we develop flexible, general processors that are relevant for large classes of AI computation and better amortize embodied carbon from semiconductor fabrication?

If successful, this research agenda will reverse current trends and permit advanced AI with lower carbon costs. Google consumes 1.5–2.3 TWh for AI, 10%–15% of its total energy use.[29] Meta attributes 30% of its AI energy to data processing, 30% to model training, and 40% to inference.[27] Studies for BLOOM's 176B-parameter language model, a GPT-3 replica, are also alarming. Training uses 433 MWh and emits 25 T-CO_2e, whereas inference uses 914 KWh and emits 19 kgs-CO_2e per day, assuming 558 requests per hour.[30]

CARBON ACCOUNTING

Research in reducing the environmental impact of AI will only be effective with the right metrics and accurate datasets. Measuring embodied carbon requires standardized methods across the industry's many companies and organizations, as well as extensible methods that accommodate new and emerging technologies. Measuring operational carbon requires scalable telemetry from large, distributed systems, such as hyperscale data centers, that track resource and power utilization. Transparency will be key to building trust and confidence.

Modeling embodied carbon from semiconductor manufacturing is difficult because complex fabrication processes are evolving to accommodate emerging technologies such as nanomaterials,[31] photonic devices,[32] and heterogeneous integration.[33] Yet, we are optimistic given recent advances in technology models and life cycle analyses.[34,35] Moreover, the manufacture of "new" technologies actually leverages many existing process flows. By mixing and matching steps in mature flows—lithography, metal and oxide deposition, etching, thermal annealing, etc.—we might estimate carbon for flows not yet in production. For example, the first monolithic 3D process flow that integrates next-generation transistors and resistive random access memory (RRAM) re-orders existing steps and adds one new step.[36]

Operational carbon depends on the energy consumed, and we need energy profilers for individual tasks, helping operators track usage and guide management. System telemetry will be combined with grid telemetry, but estimating electricity's carbon intensity is non-trivial. The marginal emission rate, which depends on recently activated generation sources, may overstate carbon because data centers often receive credits from their renewable energy investments and because grids often transfer energy across regional boundaries.

Telemetry lays the foundations for attribution, which assigns responsibility for carbon to individual pieces of computation.[37] A task's operational carbon depends on its share of data center overheads. Estimating a task's share of embodied carbon requires sophisticated analysis because tasks share servers and each task uses heterogeneous mixes of hardware. Game theory and the Shapley value may provide frameworks for fair attribution.[38]

We require reliable, harmonized, and transparent methods for carbon accounting. Data centers' energy use and emissions are verifiable by using the EPA's carbon statistics for power plants and measuring energy for hardware components. Semiconductor fabrication's energy use is more difficult to verify but could leverage published sustainability reports and datasets. Open-source models for life cycle assessment (LCA) methods would lay the foundations for improving analysis and engaging stakeholders.[39] Although computing does not yet have such foundations, the EPA and California have set standards to reduce emissions from fuels using open-source tools.[40]

ENERGY ECONOMICS

Research in computing must be cognizant of the broader societal landscape. External factors may make some solutions more practical than others or may provide opportunities to amplify or accelerate anticipated benefits. Economics and policy shape pathways to carbon-efficient computing. Governments might introduce carbon trading or incentives for low-carbon energy, while the private sector could implement offset programs, leading to renewable energy contracts and credits. DR will need sophisticated markets that price electricity at its true marginal cost, encouraging users to schedule computation accordingly. Although there is extensive literature on low-carbon policies for other industries,[41] economic analysis for computing remains relatively unexplored. Data centers, often the largest grid consumers, must understand how their net-zero operations affect other consumers and society.

Given the unpriced environmental externality of carbon,[42] one might ask if society is computing too much. What is the optimal amount of computing? Will more efficient algorithms and systems drive demand for new applications, increasing overall carbon emissions? Prior research suggests that as technology becomes more efficient, its use increases, producing rebound effects that range from 10% to 40%, reducing but not eliminating energy savings.[43] However, these effects have not been studied for computing.

We need to estimate three types of rebound effects. First, direct effects occur when lower costs increase technology use. Data centers likely exhibit strong direct effects, as more efficient processors lead to data centers with more processors. Second, indirect effects arise when lower costs increase the use of other technologies. This requires understanding the interplay between hardware components; more efficient processors may require more memory. Finally, macroeconomic effects arise when lower costs encourage new applications. Efficient processors may scale the use of large AI models for tasks like conversational bots.

CONCLUSION

Computing is at a moment of profound opportunity and promise. Emerging applications are driving unprecedented growth for

systems that offer scalable performance and environmental sustainability. Despite advances toward net-zero carbon emissions, the industry's gross energy usage continues to rise, outpacing new energy installations and renewable energy deployments. A shift toward sustainability could transform how systems are manufactured, allocated, and consumed, leading to a more responsible approach to new technologies.

As researchers establish new standards for carbon accounting, they may influence policy and legislation. An interdisciplinary community dedicated to sustainable computing is needed to train the next generation of innovators in technology, economics, and policy. Partnerships between academia and industry would accelerate the adoption of sustainable practices. Only by working together can we create holistic solutions that sustain advances in computation, revolutionizing the way we live and work for decades to come.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation Expedition in Computing (CCF-2326605, 2326606, 2326607, 2326608, 2326609, 2326610, and 2326611). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of this sponsor.

AUTHOR CONTRIBUTIONS

Conceptualization overall, B.C.L. and D.B.; conceptualization for embodied carbon, D.B., G.H., V.L., and G.-Y.W.; conceptualization for operational carbon, B.C.L., M.E., L.T.X.P., C.S., and A.W.; conceptualization for driving applications, B.P. and E.S.; conceptualization for carbon accounting, U.G., Y.Y., and M.Y.; conceptualization for energy economics, A.v.B. and B.C.L.; data science for embodied and operational carbon, D.B., G.-Y.W., G.H., and M.E.; writing – original draft, B.C.L., A.v.B., V.L., E.S., G.-Y.W., M.Y., and Y.Y.; writing – review & editing, B.C.L. and B.P.

DECLARATION OF INTERESTS

B.C.L. is a visiting/consulting scientist at Google. A.W. is a member of the advisory boards for Freeflow Ventures, Verrus, and Virtualitics.

REFERENCES

1. ITU-T (2020). Greenhouse gas emissions trajectories for the information and communication technology sector compatible with the UNFCCC Paris Agreement. https://handle.itu.int/11.1002/1000/14084.

2. Knowles, B., Widdicks, K., Blair, G., Berners-Lee, M., and Friday, A. (2022). Our house is on fire: The climate emergency and computing's responsibility. Commun. ACM *65*, 38–40. https://doi.org/10.1145/3503916.

3. Kline, D., Parshook, N., Ge, X., Brunvand, E., Melhem, R., Chrysanthis, P. K., and Jones, A.K. (2019). GreenChip: A tool for evaluating holistic sustainability of modern computing systems. Sustainable Computing: Informatics and Systems *22*, 322–332. https://doi.org/10.1016/j.suscom.2017.10.001.

4. Gupta, U., Kim, Y.G., Lee, S., Tse, J., Lee, H.H.S., Wei, G.Y., Brooks, D., and Wu, C.J. (2021). Chasing carbon: The elusive environmental footprint of computing. In Proc. International Symposium on High-Performance Computer Architecture (HPCA), pp. 854–867. https://doi.org/10.1109/HPCA51647.2021.00076.

5. Pirson, T., and Bol, D. (2021). Assessing the embodied carbon footprint of IoT edge devices with a bottom-up life-cycle approach. J. Clean. Prod. *322*, 128966. https://doi.org/10.1016/j.jclepro.2021.128966.

6. International Energy Agency. Korea 2020; Energy Policy Review. https://www.iea.org/reports/korea-2020.

7. TSMC (2025). Research Areas/Memory. https://research.tsmc.com/page/memory/4.html.

8. Gupta, U., Elgamal, M., Hills, G., Wei, G.Y., Lee, H.H.S., Brooks, D., and Wu, C.J. (2022). ACT: Designing sustainable computer systems with an architectural carbon modeling tool. In Proc. International Symposium on Computer Archiecture (ISCA), pp. 784–799. https://doi.org/10.1145/3470496.3527408.

9. Coudrain, P., Charbonnier, J., Garnier, A., Vivet, P., Vélard, R., Vinci, A., Ponthenier, F., Farcy, A., Segaud, R., Chausse, P., et al. (2019). Active interposer technology for chiplet-based advanced 3D system architectures. In Proc. IEEE Electronic Components and Technology Conference (ECTC), pp. 569–578. https://doi.org/10.1109/ECTC.2019.00092.

10. Malladi, K.T., Nothaft, F.A., Periyathambi, K., Lee, B.C., Kozyrakis, C., and Horowitz, M. (2012). Towards energy-proportional datacenter memory with mobile DRAM. In Proc. International Symposium on Computer Architecture (ISCA), pp. 37–48. https://doi.org/10.1109/ISCA.2012.6237004.

11. Semiconductor Research Corporation (2021). The decadal plan for semiconductors. https://www.src.org/about/decadal-plan/.

12. Boston Consulting Group (2024). U.S. data center power outlook: Balancing competing power consumption needs. https://www.linkedin.com/pulse/us-data-center-power-outlook-balancing-competing-consumption-lee-iz4pe/

13. Goldman Sachs Research (2024). Generational Growth: AI, data centers and the coming US power demand surge. https://www.goldmansachs.com/pdfs/insights/pages/generational-growth-ai-data-centers-and-the-coming-us-power-surge/report.pdf

14. Electric Power Research Institute (2024). Powering intelligence: Analyzing artificial intelligence and data center energy consumption. https://www.epri.com/research/products/000000003002028905.

15. SemiAnalysis (2024). AI datacenter energy dilemma – race for AI datacenter space. https://semianalysis.com/2024/03/13/ai-datacenter-energy-dilemma-race/

16. International Energy Agency (IEA) (2024). Electricity 2024: Analysis and forecast to 2026. https://iea.blob.core.windows.net/assets/6b2fd954-2017-408e-bf08-952fdd62118a/Electricity2024-Analysisandforecastto2026.pdf

17. Wierman, A., Liu, Z., Liu, I., and Mohsenian-Rad, H. (2014). Opportunities and challenges for data center demand response. In Proc. International Green Computing Conference (IGCC), pp. 1–10. https://doi.org/10.1109/IGCC.2014.7039172.

18. Acun, B., Lee, B., Kazhamiaka, F., Maeng, K., Gupta, U., Chakkaravarthy, M., Brooks, D., and Wu, C.J. (2023). Carbon Explorer: A holistic framework for designing carbon aware datacenters. In Proc. International Symposium on Computer Architecture (ISCA). https://doi.org/10.1145/3575693.3575754.

19. Xing, J., and Lee, B.C. (2024). Datacenter demand response for carbon mitigation: From concept to practicality: Invited paper. In Proc. International Green and Sustainable Computing Conference (IGSC), pp. 142–144. https://doi.org/10.1109/IGSC64514.2024.00034.

20. Fan, X., Weber, W.D., and Barroso, L.A. (2007). Power provisioning for a warehouse-scale computer. In Proc. International Symposium on Computer Architecture (ISCA), pp. 13–23. https://doi.org/10.1145/1250662.1250665.

21. Fan, S., Zahedi, S.M., and Lee, B.C. (2016). The computational sprinting game. In Proc. Architectural Support for Programming Languages and Operating Systems (ASPLOS), pp. 561–575. https://doi.org/10.1145/2872362.2872383.

22. Yeh, C., Li, V., Datta, R., Arroyo, J., Christianson, N., Zhang, C., Chen, Y., Hosseini, M., Golmohammadi, A., Shi, Y., et al. (2023). SustainGym: Reinforcement learning environments for sustainable energy systems. In Proceedings of the Thirty-Seventh International Conference on Neural Information Processing Systems, pp. 59464–59476.

23. Huang, H., Ardalani, N., Sun, A., Ke, L., Lee, H.H.S., Bhosale, S., Wu, C.J., and Lee, B. (2024). Toward efficient inference for mixture of experts. In Proceedings of the Thirty-Eighth International Conference on Neural Information Processing Systems, pp. 84033–84059.

24. Ghorbani, B., Firat, O., Freitag, M., Bapna, A., Krikun, M., Garcia, X., Chelba, C., and Cherry, C. (2021). Scaling laws for neural machine translation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2109.07740.

25. Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Association for Computational Linguistics), pp. 3645–3650. https://doi.org/10.18653/v1/P19-1355.

26. Dodge, J., Prewitt, T., Tachet des Combes, R., Odmark, E., Schwartz, R., Strubell, E., Luccioni, A.S., Smith, N.A., DeCario, N., and Buchanan, W. (2022). Measuring the carbon intensity of AI in cloud instances. In Proc. Conference on Fairness, Accountability, and Transparency (FAccT), pp. 228–239. https://doi.org/10.1145/3531146.3533234.

27. Wu, C.J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Behram, F.A., Huang, J., Bai, C., et al. (2022). Sustainable AI: Environmental implications, challenges and opportunities. Preprint at arXiv. https://doi.org/10.48550/arXiv.2111.00364.

28. Hooker, S. (2021). The hardware lottery. Commun. ACM *64*, 58–65. https://doi.org/10.1145/3467017.

29. Patterson, D., Gonzalez, J., Hölzle, U., Le, Q., Liang, C., Munguia, L.M., Rothchild, D., So, D., Texier, M., and Dean, J. (2022). The carbon footprint of machine learning training will plateau, then shrink. Preprint at arXiv. https://doi.org/10.48550/arXiv.2204.05149.

30. Luccioni, A.S., Viguier, S., and Ligozat, A.-L. (2022). Estimating the carbon footprint of BLOOM, a 176B parameter language model. Preprint at arXiv. https://doi.org/10.48550/arXiv.2211.02001.

31. Hills, G., Bardon, M.G., Doornbos, G., Yakimets, D., Schuddinck, P., Baert, R., Jang, D., Mattii, L., Sherazi, S.M.Y., Rodopoulos, D., et al. (2018). Understanding energy efficiency benefits of carbon nanotube field-effect transistors for digital VLSI. IEEE Trans. Nanotechnol. *17*, 1259–1269. https://doi.org/10.1109/TNANO.2018.2871841.

32. Wang, C., Zhang, M., Chen, X., Bertrand, M., Shams-Ansari, A., Chandrasekhar, S., Winzer, P., and Lončar, M. (2018). Integrated lithium niobate electro-optic modulators operating at CMOS-compatible voltages. Nature *562*, 101–104.

33. Lau, J.H. (2022). Recent advances and trends in advanced packaging. IEEE Trans. Compon. Packaging Manuf. Technol. *12*, 228–252. https://doi.org/10.1109/TCPMT.2022.3144461.

34. Schneider, I., Xu, H., Benecke, S., Patterson, D., Huang, K., Ranganathan, P., and Elsworth, C. (2025). Life-cycle emissions of AI hardware: A cradle-to-grave approach and generational trends. Preprint at arXiv. https://doi.org/10.48550/arXiv.2502.01671.

35. Ragnarsson, L., Rolin, C., Shamuilia, S., and Parton, E. (2022). The Green Transition of the IC Industry (White paper). https://www.imec-int.com/sites/default/files/2022-07/Whitepaper_SSTS_FINAL.pdf.

36. Srimani, T., Hills, G., Bishop, M., Lau, C., Kanhaiya, P., Ho, R., Amer, A., Chao, M., Yu, A., Wright, A., et al. (2020). Heterogeneous integration of BEOL logic and memory in a commercial foundry: Multi-tier complementary carbon nanotube logic and resistive RAM at a 130 nm node. In 2020 IEEE Symposium on VLSI Technology, pp. 1–2. https://doi.org/10.1109/VLSITechnology18217.2020.9265083.

37. Han, L., Kakadia, J., Lee, B.C., and Gupta, U. (2025). Fair-CO2: Fair attribution for cloud carbon emissions. In Proc. International Symposium on Computer Architecture (ISCA).

38. Llull, Q., Fan, S., Zahedi, S.M., and Lee, B.C. (2017). Cooper: Task colocation with cooperative games. In 2017 IEEE International Symposium on High Performance Computer Architecture (HPCA), pp. 421–432. https://doi.org/10.1109/HPCA.2017.22.

39. Sleep, S., Dadashi, Z., Chen, Y., Brandt, A.R., MacLean, H.L., and Bergerson, J.A. (2021). Improving robustness of LCA results through stakeholder engagement: A case study of emerging oil sands technologies. J. Clean. Prod. *281*, 125277. https://doi.org/10.1016/j.jclepro.2020.125277.

40. Argonne National Laboratory. GREET Model. https://greet.es.anl.gov.

41. Abrell, J., Kosch, M., and Rausch, S. (2019). Carbon abatement with renewables: Evaluating wind and solar subsidies in Germany and Spain. J. Publ. Econ. *169*, 172–202. https://doi.org/10.1016/j.jpubeco.2018.11.007.

42. Hardin, G. (1968). The tragedy of the commons. Science *162*, 1243–1248. https://doi.org/10.1126/science.162.3859.1243.

43. Gillingham, K., Rapson, D., and Wagner, G. (2016). The rebound effect and energy efficiency policy. Rev. Environ. Econ. Policy *10*, 68–88. https://doi.org/10.1093/reep/rev017.