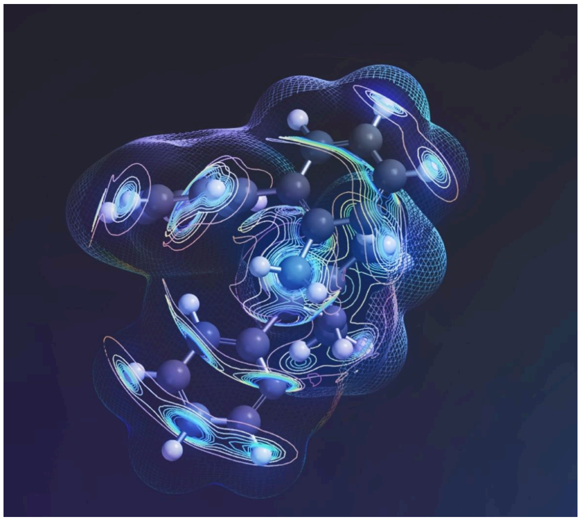
CASC | Coalition for Academic Scientific Computation

The Dynamic State of AI in Research Computing



Background

The research computing and data landscape is experiencing a profound transformation driven by surging demand to support Artificial Intelligence (AI) development and use. AI's potential to augment human capabilities is immense, and we are just beginning to unlock its powerful applications. However, this transformation comes with significant challenges. The need to build more complex AI models results in escalating demand for computing power and infrastructure. Wide use of AI algorithms in research also comes with significant social and ethical implications, as private, sensitive data is shared with AI models and algorithms for decision-making – often without transparency.

Just as high performance computing (HPC) permeated research computing and data centers decades ago, AI-enabled hardware and software are becoming a necessity rather than a luxury for the research computing and data (RCD) community. Leaders and users of RCD across the nation must be aware of the challenges and opportunities of AI to facilitate a smooth, responsible, and equitable transition to a world where AI assists discovery, innovation, and education.

Challenges

The increasing demand for running AI-powered applications means a significant shift in the operational dynamics of Research Data Centers that spans hardware, software, and human resources.

Hardware: One of the most pressing challenges is the escalating costs associated with computing for AI. Traditional data centers must be re-equipped to meet the demands of AI applications. Often, that means replacing equipment with **high-density racks of graphic processing units** (GPUs). These server racks provide computing power at a much higher density to handle the demands of AI but are power-hungry and create more heat. GPU servers often require **liquid cooling solutions**, which are much more efficient at transferring heat but are a significant financial investment. Nvidia is the main supplier of GPUs, as its proprietary CUDA platform for GPUs is dominant within the ecosystem of AI-enabled software. Although more companies are entering the AI space, **the dominance of one commercial hardware provider** can mean supply chain backups and higher prices for computing centers that already face tight budgets.

For some academic RCDs, including smaller data centers and centers at Minority Serving Institutions, the costs create a barrier to adopting cutting-edge AI-enabled hardware. These prohibitive expenses lead to unequal accessibility, widening the gap in research capabilities and hindering scientific progress. This environment could hinder efforts to **democratize AI-enabled research**, as only well-funded, large institutions can afford the necessary infrastructure. Furthermore, RCD administrators must consider the **environmental impact** of high-density, AI-enabled hardware as academic institutions strive to reconcile their growing computing needs with **organizational carbon-neutral objectives**.

Infrastructure: The integration of AI into RCDs requires storage capability that can scale as the volume grows, machine learning (ML) frameworks that support the capabilities of AI applications, high-bandwidth, low-latency networking to move massive amounts of data quickly, and robust security measures to protect sensitive data and ensure privacy. AI infrastructure must be integrated with existing legacy systems, so RCD administrators must decide whether to provide local AI resources, use the cloud AI infrastructure, or both. In addition, AI infrastructure requires scalable data storage, distributed file systems, specialized hardware with GPUs, ML libraries and frameworks like PyTorch or Tensorflow, and an extensive stack of other software packages to clean and process the data before it can be used for training or applying an AI model [1].

Workforce Development: Deploying, operating, and maintaining an AI-capable data center requires ongoing staff training. RCD staff who work with researchers must be well-versed in machine learning and deep learning concepts, AI software stacks, and GPU architectures. Because many AI advances are no more than a few years old, **there are few experts in the field**. Faculty and staff must be trained as frontline experts who can competently and effectively leverage AI technologies and assist RCD users. Training requires a **significant commitment of time and resources** and must extend beyond large institutions to create a broad, inclusive community.

The integration of AI into research computing also requires a **rethinking of RCD career paths** and job descriptions. As AI reshapes research computing, traditional roles will change, and **new roles will emerge**. RCD managers and administrators must clearly understand evolving staff roles and responsibilities and create **new career pathways** that align with future work in AI. As AI develops as a discipline, it creates uncertainty for professionals looking to enter or advance their careers in research computing. Addressing these challenges will require **strategic workforce plans** that are adaptive and forward-looking.

Security: AI systems, particularly those involved in academic research computing, handle vast amounts of sensitive data, including medical (HIPAA) data, institutional (FERPA) data, and other sensitive data sets, making them potential targets for cyber threats. Weaknesses and vulnerabilities in modern AI models create risks with characteristics that differ from traditional cybersecurity threats. These risks – including risks of attack by malicious actors and data breaches – must be considered when experts design and evaluate AI-based algorithms and systems, particularly in application domains where **trustworthiness** is essential. Meeting these critical needs for safety and security will require developing **policies for safety and security** that guide the design and evaluation of AI-based systems.

Ethics: An AI system is only as good as the data used to train it, and systems trained using biased data can lead to discrimination and unfair outcomes in everything from deciding if a business gets a bank loan to determining a medical treatment plan. Additionally, the complexity of AI systems makes it hard to understand their underlying decision-making process, and this lack of transparency and accountability can hinder efforts to identify and correct errors and biases. Finally, **the privacy of individuals** must be protected in AI-enabled research. Many widely used AI models are trained on large data sets collected from individuals by commercial companies and social media. Ensuring the confidentiality and integrity of personal and sensitive data is essential, yet hard to maintain while using AI tools. **Policies and robust data governance frameworks** must be developed to address these ethical concerns.

Actions: Enabling secure and equitable AI for research and discovery

The members of CASC support a number of actions RCDs can take to enable AI for researchers and other users.

Technologies: Research computing and data center operators, administrators, and funding organizations must understand the steps necessary to ensure RCDs can use AI securely and in ways that benefit science and society. For example, running power-hungry AI applications must be balanced with goals for environmental sustainability and cost effectiveness. GPUs are expensive to purchase and deploy. Because they generate more heat, RCDs may not be able to fulfill their environmental sustainability commitments. Adequate resources are essential to support AI implementation in RCD centers and, when appropriate, support new faculty hires who are either adept AI users or eager to learn and use AI applications. Including institutional leaders in faculty recruitment efforts could help academic institutions find new hires that can advance their AI strategies. Developing **AI computing best practices** will help with such tasks as building trusted AI data pipelines that minimize bias and return accurate query results. Guidelines must also address privacy concerns, democratized access, accountability, and social impact – and they must be updated regularly.

Education: Training and educational programs must focus on the needs of a broad audience – including technical staff, research users, students, administrators, and funding organizations. Educational efforts should support and work in collaboration with the National Science Foundation’s National Artificial Intelligence Research Resource (NAIRR) Pilot, which supports AI access and education across the nationwide research community, while gaining insights that will refine the design of a full NAIRR [2]. RCDs can leverage a number of other nationwide programs for education and training, including the NSF’s Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) [3], Strengthening the Cyberinfrastructure Professionals Ecosystem (SCIPE) [4], and EDUCAUSE [5]. Among the EDUCAUSE resources is an action plan to address gaps in AI-related policies and guidelines for AI deployment in higher education [6]. CASC also supports training programs such as the Cyberinfrastructure Leadership Academy [7], CaRCC RCD Nexus Day [8], and the annual Practice and Experience in Advanced Research Computing (PEARC) Conference [9] as avenues that will help build an educated workforce to sustain a vibrant RDC ecosystem.

Integrity and Accessibility: Faculty members, technical specialists, researchers, and students must share the responsibility for ethical AI use. This will require regular and ongoing training sessions, workshops, and seminars that articulate the potential ethical pitfalls of AI models and mitigation strategies. CASC will support the efforts of NAIRR and other groups committed to making AI accessible and usable. Diverse perspectives from all communities of users will help identify potential blind spots in AI models and applications and become feedback for the development of models that are more accurate, open, and explainable. While AI can create intelligent systems, understanding human-computer interaction will ensure that AI systems are user-centered and account for the diversity of users and uses.

Key Takeaways

The Research Computing and Data Community must act to accommodate AI. There is no stopping the AI revolution, and academic research institutions must change to keep pace with the changes AI is bringing to student learning, expectations for generative AI across the academic community, and researchers who want to use AI for data-driven problem solving. Failure to address AI in academic research computing and data centers threatens to leave the academic research community unable to fully participate in AI-enabled discovery. [10].

Training and education are essential. Resources must be committed to training programs, workshops, and other activities that allow the academic research community to become familiar with AI-enabled hardware and software. Those who work in RCDs must contribute their expertise to AI models and research problems and serve as liaisons between research teams, new AI technologies, and best practices. AI workforce qualifications must be codified, using a template such as the Campus Research Computing Consortium (CaRCC) CI Jobs Family Matrix [11]. Those working with AI must advocate for, and play a part in developing, industry standards and best practices, building on the AI Risk Management Framework developed by the National Institute of Standards and Technology [12].

AI must be balanced with energy efficiency. AI is power hungry and is becoming ubiquitous as energy costs increase and centers strive for carbon neutrality. Its benefits must be balanced with the need to manage heat emissions and create carbon-neutral data centers. Data centers will need to take advantage of chips designed to work with AI programs on GPUs and software development practices that stress energy efficiency.

Wide-ranging access and involvement are key. If the AI revolution is to be truly fair and equitable and overcome the inequalities of the past, the broad spectrum of the academic research community must be involved in AI-enabled research and education. The ability for all communities to easily access and use AI models and tools will ensure that best practices, ethical guidelines, data security, and accessibility issues address the needs of the entire research community, including smaller data centers and Minority Serving Institutions.

Authors: *Karen Green; Katia Bulekova, Boston University; Carolyn Casler, CASC; Erik Deumens, University of Florida; Jeremy Frumkin, University of Arizona; Jill Gemmill, Clemson University; Kathryn Kelley, CASC; Glen MacLachlan, George Washington University; Michael Navicky, Mississippi State University; Alana Romanella, University of Colorado-Boulder; H. Birali Runesha, University of Chicago; Semir Sarajlic, Vanderbilt University; Dan Stanzione, Texas Advanced Computing Center; Kim Wong, Pittsburgh Supercomputing Center*

About CASC

The Coalition for Academic Scientific Computation is an educational nonprofit 501(c)(3) organization with 105+ member institutions representing many of the nation's most forward-thinking universities and computing centers. CASC is dedicated to advocating for the use of the most advanced computing technology to accelerate scientific discovery for national competitiveness, global security, and economic success, as well as develop a diverse and well-prepared 21st century workforce. Learn more at http://casc.org.

References

[1] Snell, A., Olds, D., and Conway, S. (Feb. 7, 2024), HPC–AI Technology Survey 2023: Systems, CPUs, and Accelerators. Intersect360 Research.

Contact Us kelley@casc.org | (202) 670-5798 | casc.org

https://www.intersect360.com/report/hpc-ai-technology-survey-2023-systems-cpus-and-accelerators/

[2] Wright, G. (January 2022). What is Water Cooling?. TechTarget News. https://www.techtarget.com/searchdatacenter/definition/water-cooling#:~:text=Water%20cooling%2C%20also%20called%20liquid,30%20times%20faster%20than%20air.

[3] Avelar, V., et.al. (Feb. 20, 2024). The AI Disruption: Challenges and Guidance for Data Center Design. White Paper, Energy Management Research Center, Schneider Electric. https://www.itpro.com/infrastructure/data-centres/the-ai-disruption-challenges-and-guidance-for-data-center-design

[4] Bray, B.(May 20, 2023). Liquid cooling in the artificial intelligence landscape: Time to gear up. DCD Channel Blog. https://www.datacenterdynamics.com/en/opinions/liquid-cooling-in-the-artificial-intelligence-landscape-time-to-gear-up/

[5] Zhang, M. (November 5, 2023). How Much Does it Cost to Build a Data Center? Dgtl Infra. https://dgtlinfra.com/how-much-does-it-cost-to-build-a-data-center/

[6] Ananthi Claral, M.T. and Leena Rose P.J., Arul. (2019). Risks And Challenges Of Cloud Computing In Academic Field — A State–Of–Art. International Journal of Scientific and Technology Research, Vol. 8 No.12, December 19, 2019.

[7] Nlyte Software. (2024). What is an Enterprise Data Center? https://www.nlyte.com/faqs/what-is-an-enterprise-data-center/#:~:text=For%20example%2C%20you%20might%20store,the%20end%2Dusers%20it%20 serves.

[8] Hennick, C. (2016). Data Center Colocation Helps Higher Education IT Control Costs. EdTech Magazine.https://edtechmagazine.com/higher/article/2016/10/data-center-colocation-helps-higher-education-it-control-costs

[9] National Science Foundation ACCESS program. https://access-ci.org/.

[10] U.S. Department of Energy, Office of Science. https://science.osti.gov/ascr/Facilities.

[11] National Artificial Intelligence Research Resource Pilot (NAIRR), National Science Foundation. https://new.nsf.gov/focus-areas/artificial-intelligence/nairr .

[12] Kovaks, M. (Nov. 16, 2017). Lenovo launches new AI initiatives, products to boost customer productivity. Channel Daily News. https://channeldailynews.com/news/lenovo-launches-new-ai-partnerships-centres-products/57462

DOI: 10.13140/RG.2.2.30956.58244