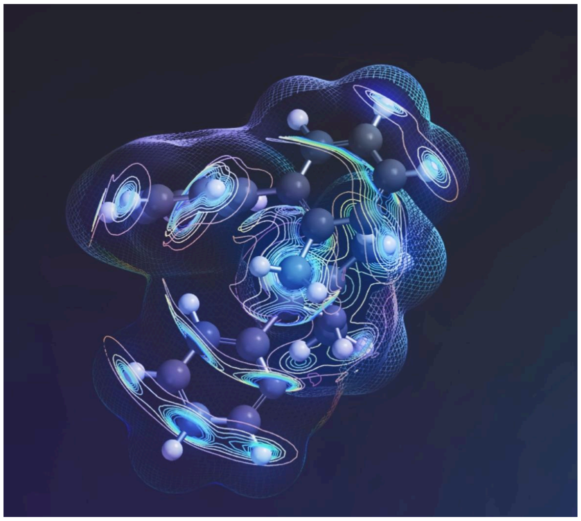
CASC | Coalition for Academic Scientific Computation

Building and Maintaining Research Data Centers in the Age of AI



Background

In this age of artificial intelligence (AI) and machine learning (ML), research computing and data (RCD) centers that provide academic researchers with high performance computing (HPC), modeling and simulation, data analysis resources and campus infrastructure must be updated or be left behind. More than 90% of academic RCD center administrators have already invested in technologies to accommodate AI–enabled applications [1]. Many are incorporating Graphic Processing Units (GPUs) – and accelerated applications that better utilize GPUs – into their computer environments because the high–bandwidth memory and parallel architecture of GPUs makes them adept at handling data–intensive tasks. To meet the demands of a

new era of computing, some organizations may want to build new RCD centers configured to handle power–hungry, densely–packed servers. Others will examine options for updating their legacy data centers as well as hybrid and cloud solutions. However they choose to move forward, RCD centers that for decades have served as the main on–ramp for scientific discovery must be adapted for a relentless pace of innovation. Institutions must make decisions about whether to host AI resources on site, in the cloud, or try a colocation solution. New research computing and data centers in the AI era require more power to process huge quantities of data and as a result generate more heat. This larger energy consumption requires innovative cooling methods such as liquid cooling, which conducts heat faster than air [2]. Many other questions related to the rapidly changing hardware and software landscapes for data must be answered if RCDs are to remain at the forefront of discovery in the age of AI.

In this paper, we outline the key challenges to building, adapting, and maintaining flexible, accessible, AI–enabled research computing and data centers and offer ideas for helping the entire research computing and data (RCD) community make a smooth transition to a new age of scientific discovery and innovation. Thousands of scientists use hardware and software at RCDs to answer research questions that affect our lives and well–being; it is critical they do not fall behind their for–profit counterparts as technology continues to evolve at lightning speed. *As an organization whose members represent RCD centers across the U.S., CASC has the community connections and the knowledge base to support efforts that help RCDs adapt, change, and flourish in this exciting new era of discovery.*

Challenges

Today's research organizations have three options to meet user needs now and in the future, all which come with their own challenges:

Evolve/upgrade an existing RCD center: Upgrading a research computing and data center to meet current and future demands requires significantly more power and resources to accommodate complicated workloads, often involving huge machine learning (ML) training data sets. Adequate power and cooling are essential to accommodate CPU servers that need more than 500 watts per CPU or GPUs that operate at more than 1,000W average per GPU card. These high–density clusters operate at 50 to 100 kilowatts per rack. While most RCDs now operate using 120/208V power distribution, some experts recommend higher voltage –such as 240/415V commonly used in the U.S. private sector or 230/400V, which has long been the standard in many countries outside the U.S. [3]. Cooling in an AI–enabled research computing and data center also requires a more efficient approach. While air cooling was the standard in data centers a decade ago, liquid cooling has become essential for AI–enabled clusters and improves performance in high–density GPU racks. Direct–to–chip (DTC) cooling, sometimes called conductive or cold plate, is currently the preferred choice and has the best compatibility with existing air-cooling systems. RCD operators and administrators will need to purchase liquid–cooled servers, which cost more but should reduce energy costs over time because of their efficiency [4].

Build a new Research Computing and Data Center: The cost of building a data center varies significantly depending on its power capacity. For a 10MW data center, construction costs typically range between $7 million and $12 million per megawatt, which translates into a total cost of $70 million to $120 million. A 5MW data center would cost between $35 million and $60 million to build, using the same per megawatt cost estimate [5]. These costs reflect the expenses associated with the building, power, cooling infrastructure, and other necessary components. The cost per megawatt can vary based on location, labor costs, and the specific requirements of the data center, such as redundancy and power density.

Whether dealing with new or refurbished facilities, RCD administrators also face the challenges of working with multiple vendors — all offering different products and services and with different commercial goals. This takes skill, time, and money and can be stressful. Academic institutions must carefully craft service agreements with vendors that make coolant distribution units (CDUs), cold plates for GPUs, networking components, and server interconnects, as well as hardware vendors to ensure all players know their responsibilities and honor their warranties.

The Cloud and Colocation: Organizations might choose to avoid the cost and expertise needed for a new data center and instead turn to the cloud. However, AI computing in the cloud can also be costly and presents its own challenges in data confidentiality, security, and management [6,7]. Moreover, cloud agreements are between a cloud provider (usually a commercial entity) and an institution. That means there is little transparency about costs that can quickly escalate as needs for data transfer and storage capacity change. Colocation is another option that lets somebody else balance power and cooling needs, however, a third party then controls physical security, including unauthorized access to hardware and preventing data loss. A lack of readily available colocation facilities and fluctuating costs also prevent some organizations from choosing this option.

Whether in a new RCD center, an upgraded facility, or a facility that offloads to the cloud, the human factors involved in RCD implementation, upgrade and maintenance must be considered. Deploying and maintaining racks of water–cooled CPU and GPU servers requires staff to utilize different skills, set up different server configurations, reconfigure networks, and accommodate changes in day–to–day operations. Operators must learn to manage hybrid centers that include services for research users who need AI–enabled HPC as well as the typical enterprise data center operations. RCD technical staff in the 2020s and beyond will need to update their technical skills, and strengthen their skills in communication, collaboration, flexibility and agility.

Actions: Envisioning diverse RCD centers with capacity and efficiency

Bringing RCD centers into the AI–enabled future will require looking at creative solutions. New centers must balance the needs of users, energy efficiency requirements, and budget realities. Costs should be evaluated not just for initial capital improvements and equipment, but for operations over a decade or more. All decision makers must understand the tradeoffs between power and capacity and the need to cool densely packed server racks.

The options to consider are:

Building a research computing and data center, which can be a challenging enterprise to embark on, as outlined in the section above. While costs can be prohibitive for some, an in–house RCD center means less reliance on third–party vendors and full control of data security protocols. A new RCD center can be built with plans to accommodate future growth, providing the scalability and flexibility to handle everything from online course offerings to cutting–edge, data–intensive research.

Retrofitting can cost millions in upgrades to mechanical, electrical and plumbing (MEP) systems, cooling capacity, monitoring and management systems, security and more. Some RCD administrators may choose to retrofit a data center if space for expansion or a new building is not an option. A retrofit can save money by keeping existing systems in use and allows RDC administrators to maintain control of their infrastructure and sensitive data. Additionally, managers must optimize their facility performance to reduce energy consumption and costs. Updating policies and procedures, such as scheduling jobs for power efficiency, can help meet users' needs for computing power. However, retrofitting is often a temporary solution. Older equipment and infrastructure, which can be updated, does not have an indefinite lifespan. Depending on the nature and extent of modifications needed, a new RCD center could end up being less costly. Investigating metrics such as power usage effectiveness (PUE), carbon usage effectiveness (CUE) and water usage effectiveness (WUE) can help retrofitted RDC centers operate more efficiently and sustainably.

Using Cloud resources can remove some of the burden of investing in AI–enabled hardware configurations. Cloud computing also means the flexibility to scale up or down as needed, although it might require a new contract with the cloud provider. However, cloud computing is expensive and means less control over infrastructure. Security, for example, is out of the hands of RCD managers, and data leaks and cyberattacks could compromise sensitive research data.

Colocation allows an organization to put server racks in another location, where someone else is responsible for temperature control, security, networking, etc. Some universities offer this on campus, allowing smaller units to locate their servers in a secure location that handles everything from campus enterprise needs to cutting–edge research using HPC. Others offer colocation services to smaller organizations to help offset their own costs or have organized with other academic institutions to share space, infrastructure, and costs [8].

Community resources can help organizations meet their research computing needs without investing in new infrastructure or cloud resources. The U.S. National Science Foundation (NSF) offers the program Advanced Cyberinfrastructure Coordination Ecosystem: Services and Support (ACCESS) to help researchers and educators with or without supporting grants to utilize the nation's advanced computing systems and services at no cost [9]. The U.S. Department of Energy's Advanced Scientific Research Division supports thousands of scientists who use HPC and advanced cyberinfrastructure through four world–class scientific user facilities [10]. The NSF has also launched the National Artificial Intelligence Research Resource (NAIRR), a pilot project that provides access to diverse AI resources [11]. These programs provide opportunities for researchers to use the most advanced computing systems and tools regardless of what resources they have locally.

Whatever actions administrators take to offer RCD users the best services possible, they must be based on a thorough review of short– and long–term funds available for building and operations, the technical talent and space available, and the resources needed to manage workloads, including enterprise workloads. The question is not simply "build vs. buy"; many options exist to make rRCD centers smarter, more efficient and more capable of handling the workloads of AI–enabled research.

CASC is well positioned to offer guidance, education and support to help RCD groups offer the best services possible to their users. For example, institutions that are new to HPC can find training and mentorship before making purchasing decisions through the Cyberinfrastructure Leadership Academy Series. CASC also supports webinars and workshops offered through Regulated Research Community of Practice (RRCoP), a community that shares costs, implementation ideas, and best practices among academic RCDs.

Key Takeaways

Whether building, retrofitting, sharing, or purchasing resources, research computing and data centers must offer the best technical solutions without sacrificing energy efficiency, affordability or accessibility. Key points to consider:

The local environment. Organizations must first look at what already exists. If, for example, a campus has a legacy data center for campus IT, can that space be upgraded to handle HPC and scientific AI applications? Is there room for expansion? Will it be possible to install water cooled systems that can conduct the heat given off by densely packed GPU servers? Will networking and security need to be upgraded? What are the budget constraints? The answers to these and other questions will determine how an organization moves forward.

Resources. Building a research computing and data center or upgrading an existing one requires a commitment of funding, personnel, dedicated space (and hopefully room to grow), and budgets for operations, maintenance and upgrades. In a university setting, these are significant expenditures, and they are ongoing, rather than one–time costs. Realistic budgets must be developed that accommodate the role of the RCDcenter now and in the future.

Efficiency. Building a research computing and data center involves meeting organizational requirements for energy efficiency. An energy audit can determine how efficiently a data center is operating prior to any upgrades. An audit can reveal the power distribution, cooling requirements, and humidity control needed to operate efficiently. An energy management plan can identify needed actions as well as the budget needed to meet efficiency goals.

Scalability. As RCD center technologies continue to evolve, so too will the needs of researchers who use them – and that means RCD administrators must scale to user needs within existing facilities or build new facilities with scalability in mind. Modularity is key so that server racks and equipment can be added as needed. Scalable data centers will need infrastructure to support advanced cooling capabilities, such as liquid cooling lines to racks and rows. The ability to expand in density, without expanding the physical footprint, will be key.

Offsite computing and shared resources. Not all RCD resources have to be local. Colocation of resources among universities can mean less downtime, shared costs, more flexibility, and better scalability. Cloud computing, although expensive, can be a solution when research groups span multiple locations and need access to large, often sensitive datasets. The key is to find a sharing solution that fits the needs of users. For example, North Carolina State University partnered with a hardware vendor for AI–enabled resources in nearby Research Triangle Park [12].

Conclusion

In order to fulfill the mission of enabling and advancing research, RCD administrators must provide the essential resources of software, CPU/GPU cycles, storage and support – all of which requires suitable infrastructure. There is no one–size–fits–all solution to building RCD centers, outside of the need for effective planning. RCD administrators and other decision makers must assess current resources, anticipate future needs and incorporate scalability for at least the next decade. Retrofitting a legacy data center should be pursued if it is feasible or necessary, although its viability will be temporary. Alternatively, constructing a new RCD center demands meticulous planning. Colocation and cloud computing may present viable alternatives, albeit costlier, and require careful consideration to avoid paying for underutilized capacity. A recommended strategy involves institutions investigating a mix of resources (on–premises/colocation, cloud, and community) to optimize cost and flexibility.

Authors: *Karen Green; Carolyn Casler, CASC; Jaime Combariza, Johns Hopkins University; Erik Deumens, University of Florida; Clark Gaylord, George Washington University; John Goodhue, Massachusetts Green High Performance Computing Center; Kevin M. Hildebrand, University of Maryland; Kathryn Kelley, CASC; Jarek Nabrzyski, Notre Dame University; John Pratt, Old Dominion University; Tabitha Samuel, University of Tennessee; J. Ray Scott, Pittsburgh Supercomputing Center; Dan Stanzione, Texas Advanced Computing Center; Bryan Webb, 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. 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

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

[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.2.2.27601.13925