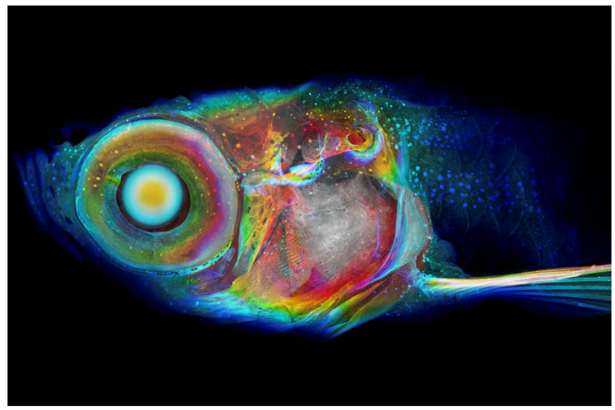
CASC | Coalition for Academic Scientific Computation

Models for Sustainability and Strategic Advancement of Institutional Research Computing and Data



Summary

Leadership in Research Computing and Data (RCD) cannot be sustained by ad hoc funding or chargeback models that fall short of cost recovery. This paper presents a framework for sustainable, resilient financial models to advance the strategic interests of the nation's research institutions.

Introduction

Excellence in research computing and data (RCD) and artificial intelligence (AI) is now central to the impact and competitiveness that all academic institutions seek in research, education, innovation, and workforce development.

The recognition that "to out-compete, we must out-compute" is fueling heavy investments in the frontiers of RCD and AI in the United States and elsewhere. However, creating and maintaining an RCD competitive edge is challenging and expensive, and becoming more so. Growing demand for infrastructure and expertise, rising costs, and the complexities of ensuring systems are compliant with evolving federal research security guidelines are threatening the capacity of many institutions to meet the moment and build for the future.

To be competitive in the fast-evolving research and technology landscape, universities must recognize research computing as a fundamental service—and adopt sustainable financial models for RCD that are driven by and aligned with institutional strategic goals, invest in people as well as systems, and seek out and build collaborations to achieve scale and maximize impact on a long-term horizon.

This paper outlines the importance of RCD as fundamental infrastructure, examines critical shifts that are challenging the status quo, and proposes a path toward sustainable and resilient financial models to maintain RCD excellence in the nation's research institutions.[1]

The Value Proposition for Investing in RCD

RCD is fundamental infrastructure for leading-edge science. Maintaining RCD excellence:

- Accelerates research productivity and scientific discovery
- Drives grant and contract competitiveness and revenue growth
- Attracts and retains top faculty, post-docs, and graduate students
- Enables interdisciplinary and data-intensive research at scale
- Improves operational efficiency for hardware and services
- Enhances institutional reputation and strategic positioning
- Positions institutions and states as leaders on national-scale priorities (e.g., AI, quantum)

The Evolving RCD Landscape and Challenge Areas

Increasing expectations on RCD centers. RCD has long been tasked with providing institutions with critical infrastructure to enable cutting-edge research and foster competitiveness in frontier science. However, most RCD centers were formed around traditional high-performance computing services. Those centers are now being asked to support both data- and compute-intensive research, driven by dynamic and interactive use cases and data-hungry AI workloads at scales not previously seen. Additionally, RCD centers are increasingly expected to connect and integrate with external resources, such as commercial clouds and an expanding array of distributed data infrastructure, to enable data streaming, edge and federated AI, and other emerging research patterns. These broadened expectations and transitions, coupled with the current rapid rise in hardware costs, are significantly increasing the cost and complexity of RCD operations and staffing.

Acquisition costs and cadence. Hardware acquisition and refresh presents one challenge. Leading-edge GPUs, high-memory nodes, and high-end networks are central to advancing

[1] A companion paper, "Regional Collaboration as Strategic Infrastructure: A Framework for the Future of Research Computing," addresses how regional collaborations provide a complementary strategy to sustain RCD capabilities beyond what individual institutions can sustain alone.

research. As GPU demand grows, costs rise, and the field is growing more complex with different domains preferring different GPU architectures. Cloud options come with higher use costs, and offer GPU architectures that are not available at most institutions. At the same time, shortages of memory and flash storage components, essential for efficiently feeding data into GPUs, are further exacerbating the financial pressures. As a result, even well-resourced institutions are struggling to keep pace with compute refresh cycles both in terms of dollars and procurement processes. What was once a non-capital purchase has become a capital purchase, delaying approvals and ultimately further increasing costs. The shift toward AI-capable systems, which are evolving rapidly in terms of specialized architecture and software requirements, is also significantly impacting RCD operating costs. The additional power, cooling, and other data center upgrades[2]; high-performing GPUs; data storage upgrades for capacity and speed; and licensing and other costs to support persistent Large Language Model services and access has put great pressure on acquisition and management/maintenance budgets.

Specialized workforce needs. People have always been an essential part of RCD infrastructure, and RCD staffing and related sustainability are not new concerns. Both hiring and retention are perennial challenges in a landscape of heavy competition from industry for the same skillsets. The National Science Foundation (NSF) has long sponsored workshops that have consistently highlighted the critical role played by the cyberinfrastructure workforce for RCD. The long-recognized gap in sustainable models for supporting the personnel who administer RCD resources has been further compounded in recent years by new market forces affecting costs and access across all facets of RCD operations, expanding demands on RCD services, and shifting funding structures, bringing the field to a transitional moment.

AI has made staffing even more critical, as RCD centers seek to acquire new types of expertise—which are in very high demand not only in the tech industry but across all industrial sectors—to support AI workflows and expanded data-related operations, and to comply with increasing federal security standards. Institutional Human Resources departments often do not have adequate sustainable staffing models or position types to respond to this specialized need for computational scientists, research software engineers, AI/ML engineers, data engineers, research facilitators, and security and compliance specialists at the level and pace needed. This has forced RCD centers to try to rely at least in part on grant funding for soft-funded RCD roles to support sustained operations, while the importance of moving beyond grant-based, soft-funded RCD roles to support sustained operations has been well-recognized [1].

---

[2] According to Turner & Townsend's annual Data Centre Construction Cost Index—the industry's primary benchmark—standard facility construction costs approximately $11.3 million per megawatt in 2026.

Budgetary and business model challenges. The above shifts are happening against a backdrop of significant budgetary constraints and shifting funding structures for many institutions. Some of the financial pressures are long-standing: The “capital cliff” that occurs when grants provide the funding for new systems but institutions lack the recurring funds to operate and replace them is not a new phenomenon. Institutions have developed a wide range of strategies to financially sustain RCD; they may centralize RCD operations, take a federated or distributed approach, outsource some portion of their RCD infrastructure, or share resources across institutions via state or regional consortia (see Appendix A for a summary of common models). The funding for these setups may be provided through campus budgets and investments (institutional subsidies or Facilities and Administrative [F&A] allocations), cost recovery (charging users), and grants or other external investments. Encountering the capital cliff has driven many institutions toward cost recovery models, but these often fall short of full cost recovery, leaving many RCD centers chronically stretched to cover operations, staff, refresh, power/cooling, long-term storage, and compliance. As a result, relying primarily on chargeback models without some level of subsidization tends to be inadequate to maintain and advance the state-of-the-art RCD services at the pace and scale needed.

Furthermore, recent uncertainties surrounding F&A costs have posed major challenges to RCD planning. Institutions with centrally funded RCD operations typically include a portion of RCD costs in the F&A rate formula; however, because the F&A rate is passed to all grants and contracts originating from an institution, this “average” approach lacks the detailed metrics for full accountability and transparency in the RCD allocation. This reliance on F&A allocations means that recent proposals to cap all F&A at 15% would bring a substantial contraction of RCD budgets in many institutions. Recognizing the need for transparency, accountability, and sustainability within the F&A cost structures, the Joint Associations Group, composed of leading U.S. higher education and research organizations, developed the Financial Accountability in Research (FAIR) model [2], which breaks indirect costs into three categories: Research Performance Costs (direct, project-specific costs), Essential Research Performance Support (project-associated support costs), and General Research Operations (broader institutional costs for research infrastructure). Whether through the FAIR model or another approach, it is clear that RCD funding structures will need to evolve to address budgetary pressures and account for the real cost of research—which includes not only the cost of doing the research, but also the cost of enabling it.

Findings and Recommendations

RCD has long played a critical role in research institutions’ impact and competitiveness, but recent developments have brought the field to a tipping point. The status quo is no longer adequate for retaining the leading edge, and for many institutions it may be barely adequate for keeping the lights on. Without sustainable RCD, research institutions are likely to find themselves increasingly burdened by aging hardware, inadequate storage, high electric bills, underutilized or inefficient systems, and shortfalls in workforce training and capacity.

We must take action now to adopt resilient financial models. These can take many forms, and should be shaped in accordance with each institution's finances, research activities, and culture (see Box 1). We offer the following recommendations to inform research funders, institutional leaders, and RCD teams as they design and implement resilient RCD models to maintain institutional excellence and retain the competitive edge of America's research enterprise.

Advancing science and innovation requires critical investments that cover the cost of conducting research as well as the computing and data infrastructure to enable it.

Recommendation: Research funders should treat research computing, data, and AI infrastructure as first-class scientific instruments [3] with explicit, sustained funding streams that cover not only capital acquisition but also operations and skilled personnel.

- Current funding mechanisms routinely assume cyberinfrastructure is either "already there" or recoverable via indirects. This is an assumption that no longer holds for AI-scale workloads, regulated data, or long-lived data stewardship.
- This is not just an NSF issue; other funding agencies including the National Institutes of Health, Department of Energy, and Department of Defense/War increasingly depend on campus RCD capacity without proportionate investment.

Research computing is a fundamental service for university research and innovation that requires sustainable funding for operations and hardware refresh.

Recommendation: Universities should fund and govern research computing and data services as institutional utilities analogous to electrical utilities, networking, libraries, and enterprise storage.

- Funding through discretionary or one-time projects is not sustainable and leads to technical debt, loss of competitive capabilities, and the eventual need for large catch-up investments.
- Utility framing aligns with researchers' expectations for RCD services and supports baseline access, equity, predictable funding, and long-term planning.
- While this cannot be achieved with cost recovery alone, treating RCD as a utility does not eliminate chargeback models entirely but suggests they are best constrained to marginal or premium services (for example, reserved or prioritized compute capacity).

Models should align incentives with institutional strategic goals and desired impact.

Recommendation: Institutional leaders and RCD teams should be explicit about what strategic goals they seek to advance and design their RCD funding and sustainability model based on the outcomes they seek to incentivize.

- Goals such as research competitiveness, broad access, long-term sustainability, and enabling growth areas can be as important as cost recovery per se.
- Models focused solely on cost recovery can limit the ability to achieve strategic goals and lead to under-utilization of resources, inequitable access, and the proliferation of fragmented stand-alone systems.
- Other RCD goals are also applicable here, such as data center consolidation, use of central storage/compute services, and cybersecurity aspirations.

Principles for RCD Sustainability Models
RCD sustainability models take many forms. The following principles are useful to inform model selection and guide strategies to support RCD resilience:

1. *Make funding mission-aligned.* Treat core RCD as a public good and anchor investment in the institution's research, teaching, and service mission.
2. *There is no one-size-fits-all.* Sustainability models should be shaped in accordance with each institution's finances, research activities, and culture.
3. *Deliver research benefits.* Models should center around serving researcher needs and improving research outcomes, transparently assess costs to researchers, support training and communication, and optimize resource utilization.
4. *Broaden access.* To support a robust research and workforce development ecosystem, institutional sustainability strategies should seek to broaden access to RCD resources and not create barriers for smaller institutions or less-funded researchers.
5. *Diversify.* Diverse funding streams can improve resilience and long-term stability by buffering against grant cycle volatility and institutional budget pressures. Diverse models can combine institutional base funding, research grant cost recovery, fee-for-service, external partnerships, industry incentives, and cloud resale margins.

The AI revolution has made RCD leadership even more central to institutional competitiveness, and regular new investment is required to maintain the cutting edge.

Recommendation: Institutional leaders should recognize the strategic opportunities of capital investment, federal funding, and industry partnerships as part of sustainability planning to achieve and maintain leadership in RCD.

- RCD infrastructure generates returns on investment by positioning institutions as regional and national leaders in innovation, enabling impactful grant-funded research, and facilitating training for the next generation of scientists and innovators.
- The cost of procurement is rising and the speed of change of RCD technologies and methods—particularly as driven by AI advancements—necessitates regular new investment to maintain the cutting edge.
- Returned overhead and chargeback models can be important elements for RCD sustainability but are not adequate to advance RCD state-of-the-art and do not preclude the need for continued capital investments.
- New capital investment should not be viewed as “zero sum” within the institution but must involve development of investment from federal, industry, philanthropic, and other external sources.
- Shared RCD infrastructure can increase the return on investment, especially for institutions with limitations in cooling capacity.

Scale matters, and competitive research computing increasingly requires scale that exceeds what many institutions can sustain independently.

Recommendation: RCD and institutional leaders should scale RCD services to the extent feasible. Those with department or college-level research computing groups should organize at the campus level, and campus-level RCD centers can achieve further scale through multi-institutional collaboration and partnerships as a cost-effective alternative to going it alone.

- Scale is not just about maximizing compute core hours or network throughput; it enhances systems and operational resilience, expertise and talent retention, security posture, regulated data compliance, vendor leverage, and operational maturity.

- There are many viable paths to scaling, including through campus, regional, consortium-based, and federated models (state systems, multi-institution facilities, shared enclaves); we endorse diverse pathways to scale, rather than a single “national facility vs. campus cluster” dichotomy.
- Collaborative models require structured interactions, extensive exchange, and clear governance, which funders, institutions, and external investors must plan for to achieve successful joint RCD partnerships.

People are a critical element of research infrastructure.

Recommendation: RCD and institutional leaders should explicitly account for people — computing, data, and networking architects; research facilitators; software engineers; data scientists; security and compliance experts; and operations personnel—as core resources in the design and implementation of RCD sustainability models.

- Workforce costs are the largest and least flexible component of RCD budgets.
- RCD specialized talent is highly sought after and difficult to recruit and retain in academic institutions. Supporting pathways to upskill existing personnel is as important as recruiting and retaining new staff.
- Grant-funded, soft-money staffing models can be important components of RCD, particularly in the innovation arena; however, overreliance on this can create systemic fragility, loss of institutional knowledge, and inequitable service.
- Hybrid funding models offer a solution in which institutions underwrite baseline staff capacity while grants and contracts support additional staffing needs.

Authors: *Bill Miller (University of Utah), Todd Shechter (University of Wisconsin–Madison), Carolyn Ellis (Arizona State University), Jaime Combariza (University of Pennsylvania), Fernanda Foertter (University of Arizona), Lindsey Fox (Vanderbilt University), Dan Majchrzak (University of Oregon), Birali Runesha (University of Chicago), Robin Scibek (Pittsburgh Supercomputing Center), Preston Smith (Purdue University), Kim Wong (University of Pittsburgh), Carolyn Casler (CASC), Kathryn Kelley (CASC)*
With the support of Anne Johnson, Creative Science Writing

About CASC

The Coalition for Academic Scientific Computation is an educational nonprofit 501(c)(3) organization with 110+ 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 an expansive, well-prepared 21st century workforce. Learn more at http://casc.org.

References

[1] National Science Foundation. (2021). Transforming Science Through Cyberinfrastructure: NSF's Blueprint for a National Cyberinfrastructure Ecosystem for Science and Engineering in the 21st Century. https://nsf-gov-resources.nsf.gov/files/CI-LWD%202.pdf.
[2] Joint Associations Group on Indirect Costs. (2025). The Financial Accountability in Research (FAIR) Model. https://www.aau.edu/key-issues/financial-accountability-research-fair-model.
[3] National Science and Technology Council, Subcommittee on Research and Development Infrastructure, Committee on Science and Technology Enterprise. (2021). National Strategic Overview for Research and Development Infrastructure. https://bidenwhitehouse.archives.gov/wp-content/uploads/2021/10/NSTC-NSO-RDI-_REV_FINAL-10-2021.pdf.

Appendix A: Common RCD Business Models

Summary of the most common frameworks for RCD sustainability.[3]

Model type	How it works	Upsides	Downsides
Centralized Core Facility	RCD is centrally funded as essential research infrastructure with minimal or no user charges (like a campus core facility or library)	Maximizes access, mitigates some budget uncertainty	Sustainability and scalability are tied to institutional finances and vulnerable to cuts
Cost-Recovery/ Recharge	RCD operates as a formal service center; users are charged according to rate structures designed to recover total cost of ownership/operations	Revenue scales with demand	Higher administrative overhead, grant funding cycles lead to volatility, can create barriers to access
Subscription/ Shared Funding/ Core Facilities	Units pay annual subscriptions for pooled access to shared infrastructure; subscription tiers may determine compute allocations, storage capacity, or queue priority	Revenue predictability	Unpredictable usage levels can lead to higher administrative overhead
Federated/ Distributed	Infrastructure is distributed across units with partial central coordination	Compute tends to be closer to the science, embedded with the science	Can lead to security issues and friction in joint projects, challenges with sharing data sets, may not take advantage of economies of scale (especially in staffing)

[3] For further exploration of RCD business models, see Lifka, D. et al. Sustainable Funding and Business Models for Academic Cyberinfrastructure Facilities Workshop II. NSF-sponsored workshop, Coalition for Academic Scientific Computation, Westin Arlington Gateway, October 2–4, 2013. https://srccii.cac.cornell.edu/

Hybrid Diversified Model	Institution subsidizes core access to baseline services; users pay extra for premium capacity and higher usage	Reduces barriers to access, diversifies revenue sources	Greater financial and governance complexity, requires careful calibration to prevent unsustainable demand, does not usually provide full cost recovery
Subsidized Condo	Users or units purchase shares of centrally-managed, institutionally-subsidized RCD infrastructure, guaranteeing priority access; resources are available to the broader community when not in use by owner	Accelerates capital acquisition	Resources may not be accessible to all researchers, can lead to equity concerns