Pushing Boundaries, Uniting Campus
Secure Research Data + Compute

Submitted By
Dr. Shawna Dark, Chief Academic Technology Officer & Executive Director | Research, Teaching, and Learning | University of California, Berkeley |shawna.dark@berkeley.edu | 510-643-9923

Project Team
- Research IT (Research, Teaching, and Learning): Chris Hoffman, Jason Christopher, Heather Amato (Public Health), James Duncan (Biostatistics), Ken Lutz, Setareh Sarrafan, Ronald Sprouse (Linguistics), Mark Yashar
- Lawrence Berkeley National Lab: Karen Fernsler, John White, George Robb, Gary Jung
- Information Services & Technology: Blaine Isbell, Gwen Davies, Joe Silva, Joleen Locanas

Project Summary

The Secure Research Data and Compute service provides secure computing and storage platforms, expert consultants, a community of practice, and the combined talents of a deeply-collaborative partnership of campus units to researchers working with highly sensitive (P4 and HIPAA) data.

The Problem
For over a decade, Research IT has partnered with researchers at UC Berkeley to identify their data and computational needs. We've seen scholars across disciplines increasingly undertake experiments, run models, and perform analysis using highly sensitive data that must be handled securely. We've watched high performance computing become a necessary tool in the life sciences and beyond, for machine learning, image analysis, and now, on the frontiers of medicine. We've also observed PIs, postdocs, and graduate students grappling with how to protect their data and design the sophisticated computing environments they need to do their work.

Unable to find appropriate support from campus, Berkeley researchers have turned to their own bespoke solutions and workflows, creating systems that are often inefficient, unsustainable, or vulnerable to attack. Their efforts sap time and energy from scientific investigation, limit the capacity for others to leverage the financial investment, and heighten the risk of data loss and data breach that would have major negative consequences for UC research and research subjects.

For Berkeley to push the boundaries of knowledge, it needed a united effort to produce secure, forward-looking infrastructure that frees researchers to focus on the questions in front of them.

The Solution

Since January 2017, the Research Data Management program — a partnership of Research IT and the Library — has collaborated with campus partners to envision, architect, build and, finally, to run the Secure Research Data and Compute service (SRDC). Bringing computation and data together in a holistic package, SRDC lets researchers safely, easily, and economically manage and work with highly sensitive data. This innovative, secure system is the product of the combined talents of staff in IT, Information Security, the D-Lab, the Office of the Vice Chancellor for Research (including the Industry Alliances Office), and Lawrence Berkeley National Lab. SRDC launched in 2020 with funding from the Vice Chancellors for Research and Undergraduate Education and the Chief Information Officer, with the support of the Chief Information Security Officer, Deans, Chairs, and researchers.

The SRDC platform offers virtual machines and high performance and high throughput computing together with secure data storage and transfer. State of the art security systems are integral to its design and implementation. The entire platform is P4 (and soon, HIPAA) compliant.

The SRDC Faculty Compute and Storage Allocation provides a baseline amount of capacity at no cost to campus researchers. A “condo” model facilitates investment in the system using grant and startup funds. Research facilitation includes access to expert consultants, documentation, training, and a growing community of practice.

First and foremost, SRDC is shaped by the needs of researchers:

- Virtual machine environments provide familiar Windows and Linux desktops to researchers not comfortable with the command line. These can be scaled to fit large jobs, running on some of the largest virtual machines ever used on campus.
- The shared high performance and high throughput computing (HPC and HTC) environment enables computationally intensive computing over sensitive datasets.
- The unified high performance file system allows researchers to work with their data across both virtual machine and HPC/HTC styles of computing.
- Data sets and software packages are tailored to research needs.
- Data ingress and egress procedures are configured to fit the project and secured in order to protect data during movement.
- Encryption is done on the fly and encryption key management, invisible to the user.

- Sandbox environments, test scripts, sample data sets, and support for containers allow researchers new to large-scale computing to test drive the systems.
- Processes and guidance woven into the fabric of SRDC help researchers manage their security responsibilities and uphold their obligations.
- Research IT consultants provide professional support for researchers seeking to use the platform. This consulting practice is deeply collaborative and relies on an extensive network of partnerships built over years. Consultants track the evolving security frameworks and the threat landscape in general, translating those requirements into actionable information and best practices for researchers.
- Close coordination among the Office for the Protection of Human Subjects, the Industry Alliances Office (which reviews and signs sponsored research and data use agreements), the Information Security Office, and the Campus Privacy Officer smoothes the campus approvals process and the acquisition of data.

SRDC crowns a growing set of campus services for researchers using sensitive data of all data classifications. Together, these new offerings expand the campus' capacity to support 21st century computational and data needs, unburden researchers, and advance research that creates knowledge, accelerates the time to discovery, and addresses the great challenges of our time.

Impact

"I've loved working with the staff at SRDC since they're clearly so committed to helping get my research off the ground. There is no way that without SRDC I'd have been able to access and easily analyze a highly sensitive government data set that's crucial to my research, and I hope this data analysis will become part of a major article or book in the future. They have also made it easy for me to work collaboratively on sensitive data with postdocs and graduate students, which has made my work go much faster and more smoothly than it otherwise would have. I'm so grateful Berkeley provides this resource to faculty!"

Rebecca Goldstein, Assistant Professor of Law

Measuring Success

The success of SRDC will be measured in a range of ways:

1. *Number of users:* SRDC, though new, has attracted significant interest from the researchers. To date, this includes 23 virtual machine engagements and nine projects interested in HPC when it launches. Growth in the number of users over time will be an important metric.
2. *Breadth of disciplines/domains of SRDC users; number and size of outreach engagements:* Research IT designs services to support the entire campus, striving to reach across the research landscape. We have designed various forms of outreach to specific research communities on campus in order to identify potential users.
3. *Overall number of SRDC-related consultations; conversion of prospects to eventual users:* The number of direct requests about SRDC shows the relevance of the service, as well as its visibility. The ability to refer researchers to SRDC represents the closing of a large gap in campus services. Conversion of interest to adoption is the gold standard. At the same time, we consider ourselves successful when we direct the researcher to the resource that works best for them, even when it is not managed by Research IT.
4. *Reduction in number of bespoke systems; appeal to researchers who would otherwise secure their own computing and storage equipment:* We have researchers using SRDC virtual machines rather than workstations of their own. SRDC is now also a service provider for Haas Research Computing at the Haas School of Business, which has decided to deprecate much of its in-house infrastructure. Bigger wins will come as research labs turn from their own legacy systems to this campus supported and secured solution.

Collaboration

SRDC brings together expertise from across the UC Berkeley landscape, organized to support our researchers.

- ***Information Services and Technology -*** *Network operations, data center build-out and operations, storage and backup services, virtual machine hosting, project management, client onboarding services*
- ***Lawrence Berkeley Lab Scientific Computing Group -*** *Network and HPC cluster architecture and systems administration, IBM Spectrum Scale parallel filesystem, data transfer facilities*
- ***Information Security Office*** *- Security reviews and assessments, security operations, CalNet and Active Directory integration support*
- ***Office of the Vice Chancellor for Research*** *- Engagement with research communities and governance support*
- ***D-Lab*** *- Community through Securing Research Data Working Group partnership and graduate student Domain Consultants*

Timeframe

- 2017: In partnership with the D-Lab and the Information Security Office, Research Data Management publishes a white paper to "assess the campus sensitive-data landscape from the point of view of researchers and research administrators; gauge the demand for services and guidance; benchmark services at peer institutions; and make a set of recommendations for future work." [1]
- 2019: Proposal for SRDC Platform approved with joint funding from the Vice Chancellor for Research, Chief Information Officer, and Vice Chancellor for Undergraduate Education. Letters of support written by three deans and one executive director.
- Fall 2020: We officially launch secure virtual machines.
- Summer 2021: Secure high performance computing debuts.

Technologies Utilized

- Computation: The SRDC virtual machine service utilizes VMWare on Dell server hardware. The HPC system is a 28-node, 560-core Dell Linux cluster equipped with a high performance Infiniband interconnect making it suitable for a wide diversity of applications. Large memory nodes facilitate computational biology and genomics research.
- Storage: A secure, high performance parallel 2.5 PB filesystem utilizing IBM Spectrum Scale Data Edition software provides data-at-rest encryption, end-to-end encryption over the network, and access logging to meet the stringent security requirements.
- Network: Multiple networks are used to segregate data flows. User access is exclusively through the use of CITRIX to prevent the unauthorized transfer of sensitive information to and from the network; there is a separate management network for administration; an infiniband network for computation; and Data Transfer nodes using Globus software with High Assurance features and a HIPAA BAA for external research data flows.
- Cybersecurity: Our defense-in-depth security approach includes intrusion detection, advanced monitoring and logging of all devices, threat and vulnerability management, endpoint security, and policy, procedures, use agreements and consulting for researchers and system administrators. Guidepoint, a professional cybersecurity company, helped develop this framework. In addition to working with our Information Security Office to develop a security plan, we engaged Trusted CI, NSF's Center of Excellence for Cybersecurity, to review our framework and provide guidance.

References

[1] Securing Research Data: A Whitepaper for UC Berkeley, July 5, 2017