FACILITIES AND OTHER RESOURCES
University of Idaho

IIDS Research Computing and Data Services

Research Computing and Data Services (RCDS) is the computational backbone for research at the University of Idaho. It provides investigators with access to reliable, state-of-the-art high performance computing (HPC) and data storage infrastructure for use in analyzing and managing large volumes of multidisciplinary research data. RCDS provides the expertise and computational tools required for processing data across all stages of the scientific data lifecycle, including raw data acquisition, modeling and analysis, sharing, dissemination, and archival. RCDS technical staff include systems administrators, scientific programmers, web developers, and data managers that collaborate with researchers to transform scientific questions into meaningful results with broader impact. Users may run jobs that use hundreds of parallel processors or require large amounts of memory and can take weeks to complete. Typical high-end projects include mathematical/computational modeling, machine learning, phylogenetic analyses, interactive data dashboards and visualizations, genome assembly, protein structure modeling, and computational physics simulations.

Overview of Facilities

RCDS infrastructure is explicitly designed to manage the complex computational and storage requirements for UI researchers with very high performance and reliability. RCDS contains an advanced mix of high-performance computing clusters, powerful servers, and reliable data storage components as well as the knowledge and technical skills required to compress years of analysis into days. Funding for RCDS infrastructure was provided from a variety of sources, including the National Institutes of Health (NIH), the National Science Foundation (NSF), the U.S. Department of Agriculture (USDA), the U.S. Department of Energy (DOE), and the M.J. Murdock Charitable Trust. RCDS equipment is housed in a modern data center provided by the University of Idaho, a second campus data center in the basement of the UI Library, and in the Collaborative Computing Center (C3) data center at the Idaho National Laboratory (INL).

Data Centers

McClure Data Center: The primary IIDS Research Computing and Data Services data center is a redesigned and renovated 1400 square-foot facility in Room 124 in McClure Hall on the University of Idaho campus. Optical fiber and copper interconnections provide 1-25 Gb/s data transfer rates

within the data center, which is connected to the multi-path 10Gb/s university backbone and from there to the Idaho Regional Optical Network (IRON) and Internet2. The McClure data center has a dedicated 80KVa UPS with three-phase power and four-forced air handlers attached to redundant university chilled water systems.

UI Library Data Center: The secondary RCDS data center resides within the basement of the UI Library. This facility is also used as the primary production data center for the central campus Information Technology Services (ITS) unit and provides the most direct uplink from the UI campus to IRON/Internet2. RCDS currently maintains several equipment racks within this shared center which houses additional RCDS data storage, virtualization, and data backup infrastructure. RCDS maintains in-rack UPS units within these Library racks, while an attached diesel generator provides indefinite emergency backup power.

INL Collaborative Computing Center (C3): RCDS also manages server and storage infrastructure within the C3 facility in Idaho Falls. The C3 is a 64,000 sq. ft state-of-the-art facility with a 200-person occupancy and a large 7,000 sq. ft data center capable of supporting up to 200 water-cooled server racks with a maximum power load of 8.5MW. RCDS manages 5 equipment racks within the C3, housing servers, storage, and network hardware. The C3 currently hosts two large DOE supercomputers: Sawtooth and Falcon, with Falcon scheduled to be transferred to academic co-management in early 2022, initially led by IIDS RCDS staff.

All three RCDS data centers have rigorous physical security and access controls. RCDS facility staff have office space in room 123 McClure Hall, 441D Life Sciences South, and room 416 in the UI Library.

Computing Systems

RCDS manages one large computer cluster for research and data analysis and modeling. Our main cluster provides over 2,500 processor cores and over 8 terabytes (TB) of system memory. The servers that comprise the cluster are interconnected with 40Gb/sec QDR (Quad Data Rate) Infiniband for inter-node communication and 1Gb/sec ethernet for management. The modular design of this cluster, primarily enclosures (blade chassis) and blade servers, makes it possible to service or upgrade components without interrupting end users. Removable and redundant fans and power supplies located at the back of the enclosures provide easy access and replacement without powering down individual systems, and each enclosure contains its own network components to maximize inter-enclosure server communication. Components include Dell M1000e blade enclosures with various blade servers, Dell rack servers, and various Supermicro servers. We have 16 cluster nodes with various NVIDIA GPU accelerators.

RCDS also maintains 12 servers (various Dell and Supermicro rack servers) that are not connected to the cluster systems for jobs that require very large, shared memory machines (such as distance-based phylogenetic analyses, genome assembly, and molecular simulations), for

software development, and for investigators who are unfamiliar with or do not require a cluster environment. The most powerful servers in this group each contain 64 cores and 1 TB (1000GB) of system memory. These powerful servers are used heavily for applications such as hybrid sequence assembly of Illumina data.

RCDS manages a rich virtualization environment for hosting 100+ virtual machines (VMs) dedicated to specific applications or research projects. These VMs run web servers, applications, databases, GIS services, models-as-a-service (MaaS), custom web Application Programming Interfaces (APIs), and more. We use VMWare ESXi 6.7 and OpenNebula hypervisors. The main VMWare environment is run on a Dell VRTX converged chassis with 4x M640 Dell PowerEdge blade servers, each with 512GB of RAM and 28/56 physical/hyperthreaded Xeon cores.

Because this scale of operation falls well outside typical University of Idaho information technology and computing services, we maintain our own support infrastructure. These include several servers for storage and authentication of user accounts (LDAP), domain name resolution (DNS), internet address assignment (DHCP) and secure connections to private networks (VPN). We also provide web and database services for online documentation and training.

Data Storage Systems

We have four distinct classes of data storage. The first group is our high-performance storage (290TB available). This storage comprises faster but more expensive disk drives and multiple control systems that are linked together through a distributed file system (Lustre) that allows us to group storage components into logical units. This makes it possible to access portions of data from multiple storage devices and aggregates data reading and writing across multiple disk drives and network connections, thereby increasing overall performance. Metadata servers contain typical file system information such as ownership permissions, and physical location. We have multiple metadata servers working in parallel to recognize failures and automate device control to minimize staff intervention and disruption of services. Each individual disk storage system (array) combines multiple disks into a single logical unit (RAID), which provides redundancy on a disk level. Components currently include Dell MD3420 storage arrays, Dell R515, R510, R630 servers, and various Supermicro servers.

The second group is our commodity storage (1.9PB gross). This storage group uses cheaper, but slower, disks to house most user data. We currently run a Ceph distributed file system, which offers increased performance and reliability. Components currently include various Dell and Supermicro rack servers.

The third storage group comprises our backup storage systems (898TB gross, 630TB of which is off-site). We back up user data regularly to an offsite location on HDDs using ZFS snapshots. Components include Storinator Storage Pods and a Supermicro rack server.

Finally, our main virtualization and hosting infrastructure currently uses a 30TB ultra-high-performance, in-chassis SSD RAID-5 array to host the virtual disks for production VMs. 500TB of slower production data storage is provided by a NetApp FAS 2554 server connected to the VRTX chassis at 10Gbps. A second 500TB NetApp FAS 2554 is located within the C3 data center in Idaho Falls, primarily for the purposes of off-site backup and Disaster Recovery (DR) for our virtualization services.

Data Management and Scientific Programming Services

RCDS provides research data management infrastructure and services that enable UI investigators to store, catalog, disseminate, and archive their research data outputs. RCDS helps researchers write comprehensive Data Management Plans (DMPs) to include in their funding proposals. RCDS operates the UI's official research data repository and is actively engaged with related national initiatives such as the Data Observation Network for Earth (DataONE) and the Earth Science Information Partners (ESIP). Through our involvement with DataONE, RCDS is part of a federation of similar repositories at a global scale, thereby increasing the exposure, resiliency, and discoverability of RCDS-published research data. We can also provision Digital Object Identifiers (DOIs) for datasets through our institutional DataCite membership.

RCDS provides custom scientific programming services to help investigators design, develop, and host custom research applications and tools. These tools often include databases, web applications, and interactive geospatial mapping and data visualization. RCDS is increasingly involved in helping researchers with computational modeling, machine learning, and mobile application development.

Science DMZ

Working with the UI ITS Networking team, we have set up a Science DMZ network to allow for unfettered, high-throughput access to research data. This Science DMZ network hosts a Globus Data Transfer Node (DTN) and perfSONAR servers. Globus allows for reliable, high-speed data transfers across the Internet2 backbone, making it possible to connect safely and efficiently to computational cores at collaborators' institutions to share data. The perfSONAR servers provide a managed mechanism to test and improve latency and network throughput to other institutions.