

FACILITIES AND OTHER RESOURCES

University of Idaho

IBEST Computational Resources Core

The IBEST Computational Resources Core (CRC) serves as the computational backbone of evolutionary and computational biology research at the University of Idaho. It provides investigators with reliable and highly available state of the art high performance computing and large data storage capabilities for use in analyzing and managing large volumes of research data. We provide the computational tools required for processing data across all stages of analysis including the earliest stages of processing raw data generated by various sequencing platforms to the genomic, metagenomic, and phylogenetic analysis required to transform biological questions into meaningful results. Users of the core run jobs that may use hundreds of processors in parallel or large memory allocations and may run require weeks to complete. Typical high-end projects include mathematical modeling, machine learning, phylogenetic analyses, genome assembly, protein structure modeling, and computational biology simulations.

Overview of facilities

The CRC is explicitly designed to manage the complex computational and storage requirements for the IBEST researchers and core facilities with very high data reliability and availability. The core 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. The National Institutes of Health and the Murdock foundation have provided funding for the core. All equipment is housed in a state of the art data center provided by the University of Idaho.

Data Center

The IBEST Computational Resources Core data center is a 1400 square foot facility in Room 124 in McClure Hall on the University of Idaho campus that has been specifically designed and renovated for the core. Optical fiber and copper interconnections provide 1-25 Gb/s data transfer rates within the core, which is connected to the 10Gb/s university backbone and from there to Internet2. This room has a dedicated 80KVa UPS with three-phase power and four-forced air handlers attached to redundant university chilled water systems. Core facility staff has office space in room 123 McClure Hall and 441D Life Sciences South.

Computing Systems

We have one large computer cluster for research and genomic data analysis. Our main cluster provides over 1500 processor cores and over 8 terabytes of system memory. The servers that comprise the cluster are connected 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 our users. Removable and redundant fans and power supplies located at the back of the enclosure 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 R730 and R815 rack servers, and Supermicro 5018D-FN4T servers, and 6 Supermicro GPU servers.

We also maintain 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 terabyte (1000GB) of system memory. These powerful servers used heavily for hybrid sequence assembly Illumina data.

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.

Data storage systems

We have three distinct classes of data storage. The first group is our high performance storage (200TB 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 in order 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 Supermicro 5029P servers.

The second group is our commodity storage (1.9PB gross). This storage group uses cheaper, but slower, disks to house the majority of 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 commodity disk using ZFS snapshots. Components include Storinator Storage Pods and a Supermicro rack server.

Science DMZ

Working with the UI Networking team, we have set up a Globus data transfer node on the university's science DMZ network – which allows for high-speed data transfers on the Internet2 backbone. These servers make it possible to connect safely to computational cores at collaborators' institutions and share data

Classroom for education and training

To support educational programs and inter-institutional collaborations we maintain three teleconferencing enabled conference rooms and a state of the art technology classroom. The classroom contains 15 iMac computers purchased in 2016 and 12 iMac computers purchased in 2012 and a lectern connected to an HD projector. The classroom is used extensively by instructors from the College of Science and the College of Natural Resources. The classroom also has a Tandberg teleconferencing system, which allows us to offer workshops and classes to and from collaborating institutions. We also maintain a 16 node cluster built into a half-height rack designed to enable high performance computing administration and usage classes, and experimentation with emerging technologies.