

1 Facilities

Carnegie Mellon University has access to a range of computing facilities to run simulations, design space evaluations and prototype software. Researchers have access to a range of computing platforms: high throughput computing (1,000+ cores) for design space evaluation via HTCCondor, Slurm and commercial cloud providers, capability computing via the ECE Department's DataScience & Engineering SuperCloud (large memory machines and GPU accelerated machines) and the Pittsburgh Supercomputing Center (the *bridges* machines at PSC, huge memory and large number of CPU/GPU systems), and access to hosting of advanced FPGA compute systems (BittWare TeraBox with multiple Arria and Stratix 10 FPGAs) in the ECE department's computing lab.

1.1 CMU's computing services

The ECE department maintains central computing clusters and a high throughput HTCCondor pool providing 1000+ CPUs with mutiple terabytes of memory to faculty and students. The Carnegie Mellon intranet is a fully-interconnected, multimedia, multi-protocol infrastructure spanning hundreds of separate VLANS (virtual LANS). These segments are attached to a fully redundant, ring shaped backbone, enabling access between all systems on the campus, including the Supercomputer facilities operated by the Pittsburgh Supercomputing Center. Other research institutions that participate in the Internet2 community are also readily accessible via high-speed links . Content is delivered via 100Gbs (gigabit/second), 10Gbs (gigabit/second), 1Gbs (gigabit/second) or 100Mbs wired Ethernet or via Wireless Access Points.

1.2 Server hosting

For researchers who need to own their own dedicated compute cluster, the campus provides server hosting in a professionally managed data center on site. This allows teams to host advanced compute systems in the ECE department's computing center. It is encouraged to allow overflow sharing via HTCCondor but exclusive ownership is possible. The goal is that machines too big to be put under a desk should be housed in the data center.

1.3 DataScience & Engineering SuperCloud

The ECE department has built up a data science computing infrastructure that gives individual researchers access to powerful computing equipment

and scaled-out storage. Commercial off the shelf (COTS) multi-core machines with large amount of memory and disk space are installed with commonly used analytics software (e.g., Matlab and Python) and made available to faculty for research purposes.

1.4 Commercial cloud computing

ECE researchers may want to utilize commercial cloud computing services like Amazon Web Services (AWS) and Windows Azure. IT Services personnel are able to assist with adapting and deploying high-level architectures to cloud machines. Academic/educational pricing is available from the vendors.

1.5 Pittsburgh Supercomputing Center (PSC)

CMU faculty can gain access to resources at the Pittsburgh Supercomputing Center, and to the national computing infrastructure like NSF's XSEDE. The 800+ node *Bridges* supercomputer at the PSC is designed for Big Data users, and supports both traditional and non-traditional HPC uses. On multiple nodes, there are up to 12TB of RAM and sixteen 18-core CPUs, and Bridges features numerous 2 GPU-nodes. ECE IT services and PSC personnel assist researchers in getting access and started on using the high performance (HPC) and supercomputing resources available for data sciences and graph analytics.

Given the complexity of using true high throughput computing environments, educated and experienced users are required for such machines. IT Services and PSC Personnel can help researchers with getting access and getting started. Accessing such machines is to be handled on a case-by-case basis.

1.6 Secure and private facilities

Parts of the network are restricted to campus only use, allowing for enhanced security and/or privacy. Research facilities are managed with single-sign-on accounts and two-factor authentication. Provisioning and deprovisioning occurs automatically with changes in account status. Role-based access controls are utilized at the application, network, storage, and operating system layers. ITAR and EAR spaces are available for more restricted data. The network perimeter and internal networks are scanned regularly for system vulnerabilities. Logging and monitoring facilities operate continuously and can be analyzed with appropriate authorization.

2 Data Management

2.1 Roles and Responsibilities

[PIs should outline the rights and obligations of all parties as to their roles and responsibilities in the management and retention of research data. They must also consider changes to roles and responsibilities that will occur should a principal investigator or co-PI leave the institution.]

2.1.1 Computing Facilities Roles and Responsibilities

University Computing provides standard data classifications and guidelines for storage and security. Common storage areas with high availability, redundancy, and safety are provided to the university community for each of these areas. They include SaaS cloud storage, Andrew File System (AFS), DFS/CIFS/SMB, and local file systems attached to compute nodes. The ECE Department IT Services provides supplemental data storage locations within a managed computing environment where research is conducted, including AFS, DFS/CIFS/SMB, and local file systems attached to compute nodes.

2.2 Expected Data

[PIs should describe the types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project. It should then describe the expected types of data to be retained.]

2.3 Period of data retention

[PIs should describe the period of data retention. Minimum data retention of research data is three years after conclusion of the award or three years after public release, whichever is later. Public release of data should be at the earliest reasonable time. A reasonable standard of timeliness is to make the data accessible immediately after publication, where submission for publication is also expected to be timely. Exceptions requiring longer retention periods may occur when data supports patents, when questions arise from inquiries or investigations with respect to research, or when a student is involved, requiring data to be retained a timely period after the degree is awarded. Research data that support patents should be retained for the entire term of the patent. Longer retention periods may also be necessary

when data represents a large collection that is widely useful to the research community. For example, special circumstances arise from the collection and analysis of large, longitudinal data sets that may require retention for more than three years. Project data-retention and data-sharing policies should account for these needs.]

2.3.1 Available methods and duration of data retention

University-provided locations have short-term disaster recovery and data-retention capabilities (less than 2 months). Specific storage locations provided by the ECE Department IT Services are capable of retaining data indefinitely assuming the data is static. Project volumes in Andrew File System (AFS) and DFS/CIFS/SMB are backed up to long-term cold storage using physical or virtual linear tape. Near-line storage is used for disaster recovery of specific locations. It is the responsibility of the PI to ensure these locations are used to meet data retention requirements.

2.4 Data formats and metadata

[PIs should describe the specific data formats, media, including any metadata.]

2.5 Data dissemination and policies for public access, sharing and publication delays

[PIs should clearly articulate how sharing of primary data is to be implemented. Describe dissemination approaches that will be used to make data available to others. Policies for public access and sharing should be described, including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements. Research centers and major partnerships with industry or other user communities must also address how data are to be shared and managed with partners, center members, and other major stakeholders. Publication delay policies (if applicable) must be clearly stated. Investigators are expected to submit significant findings for publication quickly that are consistent with the publication delay obligations of key partners, such as industrial members of a research center.]

2.5.1 Facilities for access and sharing

The University and the ECE Department IT Services provide web site locations accessible to the public. Data not ready for public consumption is stored on systems with role-based access control restrictions and secured within on-campus locations or is stored in a SaaS provider service using strong encryption in transit and at rest. Cryptographic keys are withheld from SaaS providers to prevent unapproved decryption. Data classified by the University as private or restricted must be handled in adherence to the established guidelines. Accounts used to access data are provisioned and deprovisioned automatically as account status changes. Two-factor authentication and single-sign-on for accounts is commonly available to the University community.

2.6 Data storage and preservation of access

[PIs should describe physical and cyber resources and facilities that will be used for the effective preservation and storage of research data. In collaborative proposals or proposals involving sub-awards, the lead PI is responsible for assuring data storage and access.]

2.6.1 Facilities for preservation and storage

University Computing provides common storage areas with high availability, redundancy, and safety. They include SaaS cloud storage, Andrew File System (AFS), DFS/CIFS/SMB, and local file systems attached to compute nodes. The ECE Department IT Services provides supplemental data storage locations within a managed computing environment where research is conducted, including AFS, DFS/CIFS/SMB, and local file systems attached to compute nodes. University-provided locations have short-term disaster recovery and data-retention capabilities (less than 2 months). Specific storage locations provided by the ECE Department IT Services are capable of retaining data indefinitely assuming the data is static. Project volumes in Andrew File System (AFS) and DFS/CIFS/SMB are backed up to long-term cold storage using physical or virtual linear tape. Near-line storage is used for disaster recovery of specific locations. It is the responsibility of the PI to ensure these locations are used to meet data retention requirements.