University of Florida Campus Cyberinfrastructure Plan 2015-2020

Executive summary

The University of Florida formulated a Campus Cyberinfrastructure plan in 2011 for 2011-2015. The accomplishments of that plan are listed below. This plan lays out the goals for the next five years 2016-2020. The needs as perceived by the faculty and their research associates list the goals for the next 5 years as follows:

- Flexible network infrastructure to optimally make use of the investments in infrastructure (UFNet2).
- Support for research projects that need to handle restricted data such that contracts that include regulatory and legal requirements, such as HIPAA and FISMA, do not pose insurmountable obstacles (GatorVault).
- Provide advanced services for research projects like ready-to-run applications, and persistent databases and web portals, such that the research project does not have to take on system administrator talent and duties.
- Develop a framework for data lifecycle management policies, guidelines and standards to be prepared for increased requirements from funding sources.
- Sustain growth by expanding data storage and computing capacity, and network bandwidth to match the demand. Increase the capabilities offered to support multi-institution research projects that work with big data. Implement IPv6 for worldwide projects like CMS Tier-2.

Accomplishments

Built rich infrastructure: The University of Florida offers its faculty a balanced portfolio of compute and storage options: on-campus services and using national resources like XSEDE, all supported seamlessly by on-campus experts, with an in-person training program that also offers online resources.

o **Provide physical space:** The 25,000 sq. ft. 1.75 MW data center was completed in January 2013. It provides 5,000 sq. ft. of space for research computing and 5,000 sq. ft. for general enterprise computing services, including teaching and enterprise application support.
o **Improved network infrastructure:** Upgrade of the existing 20 Gb/s Campus Research Network to 200 Gb/s was completed on January 31, 2013 (with funding by NSF MRI ACI-1229575 award). At the same time the connection to Florida Lambda Rail (FLR) was upgraded from 10 Gb/s to 100 Gb/s and the FLR link to Jacksonville from 10 Gb/s to 100 Gb/s (with partial funding by an NSF CC-NIE ACI-1245880 award). UF joined the Internet2 Innovation Platform[1] leveraging its 2004 investment in a Science DMZ and work by its faculty on software defined networking (SDN). Recent records show that the 100 Gb/s link was used by multiple institutions to support sustained 40 Gb/s data transfers. The Florida LambdaRail network is upgrading its backbone from 10 Gb/s to 100 Gb/s by June 2015. The

---

[1] Announcement http://internet2.edu/news/pr/2012.04.25.disruptive-tech-to-advance-us-scientific-research.html

UF campus offers Eduroam to allow academic visitors to connect to the wireless with home-institution credentials.

o **Increased compute and storage capacity:** HiPerGator, a supercomputer with 16,000 new cores, started production in August 2013, bringing the total available core count over 21,000. The research data storage capacity to 5 PB. Over 15,000 cores have been sold to research projects to provide guaranteed access in addition to the possibility to use idle cores for free when they are available.

Services and expertise UF built a set of well-defined compute and storage offerings with expert consulting services and training to make researchers more efficient, productive, and competitive in obtaining funding for their proposed research projects. Established a common strategy for data-projects across campus in collaboration between Research Computing and the Libraries.

Collaborative research storage With other universities in the State of Florida and partner DataDirect Networks, a system was built[2] that allows researchers to store data in ways that can be accessed locally from desktops and campus supercomputers. The system then replicates data to participating institutions as controlled by data policies so that collaborators can access the data locally to them on their campus resources.

Background

In April 2011 the University of Florida created coherent organization to support the computing needs of researchers. The organization is called UF Research Computing[3] and reports to the Office of IT with major additional funding from the Provost and the VP for Research. The new organization has built on the success of the UF HPC Center that was established in 2005 as a collaborative effort between faculty, departments, colleges and the Office of IT. UF Research Computing is governed by the Research Computing Advisory Committee[4] which sets strategic direction and policy. The committee has actively reached out to include the digital humanities and social sciences, in collaboration with various groups including the UF Digital Humanities Working Group which began in 2011. The Data Management/Curation Task Force[5] has been exploring the possibilities to coordinate data life cycle management training and resource development since 2013. In 2014 the Geospatial Task Force[6] started to work on a strategic plan to optimize the use, training, support, and dissemination of geospatial thinking in education and research across the campus. In 2014 the Informatics Institute[7] was created to foster and coordinate interdisciplinary research and collaboration around data and information. In 2014, the UF Digital Humanities Graduate Certificate Working Group was created to support graduate education in the Digital Humanities, in part drawing on the resources, training, and expertise provided by Research Computing in collaboration with the Libraries. In 2015, the statewide Florida Digital Humanities Consortium began and one of the goals is to further enable Digital

---

[2] Announcement http://news.it.ufl.edu/research/new-storage-system-simplifies-bigdata-sharing/
[3] Website http://www.rc.ufl.edu
[4] Website http://www.it.ufl.edu/governance/advisorycommittees/researchcomputing.html
[5] Website http://cms.uflib.ufl.edu/datamgmt/index/TFCharge.aspx
[6] Website http://www.it.ufl.edu/governance/advisory-committees/geospatial-task-force/
[7] Website https://informatics.research.ufl.edu/

Humanities activities, in part through collaboration with research computing entities across institutions.

All scientific, engineering, scholarly, and educational activities are deeply impacted by developments in information technologies. These innovations improve upon established approaches as well as make it feasible to follow previously unexplored paths to manipulate and investigate 'big data' in multiple ways. To ensure that its researchers, scholars, and students remain productive and competitive in the coming decade, a university must provide the necessary framework and tools.

The first five-year Campus CI plan 2011-2015 provided the basic infrastructure and the deployment and adoption by a large fraction of the campus community of a set of basic services, and support and training offerings. The focus of these efforts was very strongly driven by faculty demand, as expressed through committees, surveys, and many person-to-person discussions and interviews.

It has become clear that to allow advanced researchers to move to the next level, specific enhancements and added services are needed. Furthermore, to be able to support the next class of researchers in their data analytics and computing needs, further refinements and enhancements of services are needed. This Campus CI plan 2015-2020 outlines the strategy to address these needs.

Goals for campus cyberinfrastructure

Sustain growth to meet demand

To keep up with growing demand for computing and data storage needs, the next system to be added to HiPerGator, called HiPerGator2, is being designed. The target date for production is August 2015. The existing 2 PB storage system of HiPerGator2 will be expanded to 3 PB. The infrastructure is partially funded by the University and partially through cost recovery from grants and contracts. This has proven to be a long-term sustainable model with over 5 years of data and experience. Limited access is provided for free to avoid barriers to adoption.

The collaborative research data cluster built by SSERCA (Sunshine State Education and Research Computing Alliance) now comprises three institutions (FSU, UF, and USF), with several others planning to join in the coming year. This infrastructure makes collaborative research teams comprised of faculty at institutions in the State of Florida more competitive. The current storage pool is about 500TB. As more sites join and demand grows, the pool will be expanded. The business model across the State is being developed and will align with the practices on each campus. The service is free to support new users within a small limit and is charged to grants and contracts for larger storage needs and higher service level expectations.

Virtual network environments: UFNet2

The campus has invested in top-tier infrastructure for storage, computing, and networking consolidated in the new data center. To ensure that every researcher can effectively and efficiently use this infrastructure, in the style of a private-cloud, investments are being made to

provide fast and flexible network infrastructure to the desktop, laboratory, and mobile devices. The number of wireless access points has been increased and service to network ports will be at 1 Gb/s by the end of 2015.

In addition, the backbone has been upgraded to support the virtual routing and forwarding (VRF) stack. This will allow the deployment of several virtual networks across all enabled buildings. Currently the network security and policy is dictated on a per building basis by the most restricted data present in the building. As a result, all researchers located in buildings that have clinical work and research going on, and thus have protected health information (PHI) in them, must work in the restricted network environment, whether they actually touch PHI or not. The new UFNet2 infrastructure will allow academic, health, and science DMZ network environments to be deployed to any port depending on how the device is registered. Thus computers of researchers not engaged with PHI data can be moved to the more open academic network environment; gene sequencing instruments can be moved directly to the research science DMZ network environment. The deployment timeline for the buildings is being developed and will extend through 2015 and 2016.

Advanced services for research: GatorCloud

The computational power of HiPerGator and its storage capacity for large data projects can be accessed by users who are familiar with, or are willing to become familiar with, the traditional high-performance computing way of working: connect with ssh to a command line prompt.

There is a large group of researchers who are dealing with data sets that are increasing in size to grow beyond the capability of their desktops to handle. It is not an efficient use of the time of these researchers to become command-line HPC users. In addition, the software they use often exists only in the Windows version, or the Windows version has much better support and more features than the Linux version. It is possible to provide these users with applications that run in virtual machines---with a Windows OS as needed---running on the shared HiPerGator infrastructure with access to the fast data storage systems. The infrastructure is being deployed in 2015 and the services will be packaged under the collective name of GatorCloud.

In addition to specific applications like MATLAB, SAS, R, it will also be possible for programmers and developers to provision virtual machines and clusters of virtual machines for testing and deployment of services that need the large data storage and computing capacity of HiPerGator.

Increasingly funded research projects have the requirement to actively share data with collaborators, both during the funded life of the project and for some time after the project is completed and funding has ended. These requirements take on the form of persistent databases, maybe with an authenticated-web-page front end, or of a fully featured portal that allows collaborators to work on project data during the project and supports general users to browse published data after the project has completed. The functionality of the portal may include allowing the authenticated user to upload their own data set, run the project's published algorithms using back-end computational resources like HiPerGator, and then download the results.

Developers in the project need to be able to develop the algorithms, collaborators need to be able to enter data into the database, but there should not be a requirement for the project to have in-house system administrator skills to ensure that the webservers, databases, and portals are functional, secure, and properly maintained.

Infrastructure for work with restricted data: GatorVault

The data storage, computing, and networking infrastructure available today is limited to use in research projects that work only with open data. There is a significant need to have a uniform, well-documented, well-supported infrastructure for work with restricted data, in particular protected health information (PHI). The planning for this project started in 2013 and it will go into production in multiple phases, with FISMA compliance in place by June 30, 2015. Further features will be added in 2016.

The architecture will be of a data enclave---a vault---with the ability to store data and to process the data inside the enclave using virtual desktop technology. Thus data will not be transferred out to user end-point devices, which will be one of the elements reducing the risk for data compromises. The system has been designed to have sufficient computational capacity so that modern high-performance data analytics approaches can be carried out inside the data vault.

The requirements for the advanced, persistent services in support of research are not limited to open data, or data to be protected for intellectual property, but also include restricted data that must be handled in ways that meet regulatory and legal requirements.

Data life cycle policies, guidelines, and standards

As data and collaboration become more important it is expected that guidelines from funding agencies will become more detailed and may turn into requirements before 2020. The Data management/Curation Task Force is developing a framework for discussing the creation of policies, guidelines, and standards for data life cycle management on campus to prepare for such requirements.

Implementation of IPv6

The infrastructure to support IPv6 protocols has been planned and critical services such as routing and name service (DNS) have been implemented on test systems and networks. The goal is that the Science DMZ will carry live IPv6 traffic to public Research Computing resources by summer of 2015: webserver, HPC login and submit node, data transfer nodes running the Globus server. In addition, the compute nodes of HiPerGator2 will operate in dual-stack mode so that the workflows for the CMS experiment can return result-data to the home institutions of the high-energy physicists using the UF Tier-2 services without the use of the network address translation (NAT) protocol.

InCommon authentication

The primary system for campus-wide authentication is the GatorLink Shibboleth infrastructure with identities issued to UF students, faculty, staff, as well as alumni and various associated. During 2012, UF has made infrastructure improvements to achieve InCommon Silver certification for specific subsets of users, such as those who use two-factor authentication.