On the Development of a Next-Generation, Federated Cyberinfrastructure: NCAR’s Perspective

The Computational and Information Systems Laboratory (CISL) at the National Center for Atmospheric Research (NCAR) is focused on pursuing a cyberinfrastructure federation strategy that aligns well with its mission of providing discipline-specific computing and data services to the atmospheric and related sciences community, on the one hand, and its recognition that this mission carries well beyond our own facilities and capabilities to include peer organizations both nationally and internationally. In the sections below, we first describe CISL’s current activities in fostering CI federation and development, and in the second section discuss a vision for fostering future federated CI. Finally, we touch on a few of the challenges we see in federated cyberinfrastructure in the future.

NCAR/CISL’s relevant CI activities

**NCAR as part of XSEDE.** CISL currently participates in the XSEDE federation as a “Level 2” Service Provider, with integration focused on maintaining connectivity via XSEDEnet and using Globus data transfer services to integrate with other XSEDE service providers. We believe that it is a good thing that the vision of federated CI has evolved to be much broader in XSEDE than the earlier TeraGrid approach, with its focus on tight compute resource integration, to encompass a broader set of services, many of which are data-centric in nature.

Beyond XSEDE, CISL is actively engaged with a variety of organizations and in a broader spectrum of CI federation activities at the regional, national and international levels. CISL is also keen to ensure that its CI is fully integrated with emerging standards and technologies.

**Other National CI Engagement.** CISL is a recognized leader and participant in a number of national networking projects that are tightly integrated with regional networks. CISL’s involvement with national networking consortia includes Internet2 (I2) and the Department of Energy (DOE) Energy Sciences Network (ESnet). In addition to networking, NCAR plays a leadership role in national data-related activities including the Advanced Cooperative Arctic Data and Information Service (ACADIS), the North American Regional Climate Change and Assessment Program (NARCCAP), the Community Data Portal (CDP), and the Chronopolis digital preservation system.

**Regional CI Engagement.** CISL is a recognized leader and participant in many regional networking projects that are tightly integrated with national networks. CISL’s involvement with regional networking consortia includes FRGP, BiSON, WRN, EAGLE-Net, CHECO, BRAN, the Colorado Broadband Roundtable for the IT Economic Development and Business Planning group with the Colorado Governor’s Office, and the Quilt. These are all designed to provide NCAR/UCAR and other institutions in the region with robust wide-area data pathways. CISL designs, plans, engineers, installs, operates, maintains, develops strategy, and performs research for NCAR and UCAR state-of-the-art data networking and telecommunications.
facilities, providing a vital service to NCAR, UCAR, and UCP research communities by linking scientists at partner institutions throughout the western U.S. to NCAR’s supercomputing resources.

CISL also participates in the activities of the Rocky Mountain Advanced Computing Consortium (RMACC), a consortium of regional HPC organizations including Colorado State University, the University of Colorado, Colorado School of Mines, the University of Wyoming, NCAR, the Department of Energy’s National Renewable Energy Laboratory, and NOAA’s Boulder Earth System Research Laboratory. RMACC representatives meet quarterly to discuss and organize joint training, education, and outreach activities.

**International CI Engagement.** CISL is part of several noteworthy large data federation efforts. It is part of the Earth System Grid Federation (ESGF) Peer-to-Peer (P2P) enterprise system, a collaboration that develops, deploys and maintains software infrastructure for the management, dissemination, and analysis of model output and observational data related to the question of climate change.

CISL also contributes to developing the World Meteorological Organization (WMO) Information System (WIS). Under the auspices of the United Nations, the WMO is designing, developing, and deploying WIS as a next-generation globally federated information system for weather, climate, hydrology, oceanography, and many other disciplines.

**Authentication and identity management.** CISL has also pursued opportunities to better integrate its infrastructure in terms of identity management and authentication. CISL is a TAGPMA-accredited certificate authority, which allows CISL-issued X.509 certificates to be recognized by other sites. CISL has joined the eduroam federated network access service, which allows visitors to immediately and securely join UCAR’s wireless network. CISL is also using InCommon partially to participate in IGTF and therefore XSEDE and Globus.

**CISL’s vision for future federated CI development**

**It’s all about the data.** Atmospheric science is data-intensive and CISL has placed significant emphasis on developing a data-centric approach to serving our community. As part of the ESGF, for example, CISL serves over 140 terabytes through this system to the international climate science community. The CISL-operated Research Data Archive curates 600 data sets representing 1.5 PB of data that support scientific studies in climate, weather, Earth System modeling, and increasingly, other related sciences. CISL is a community leader in developing data services that address the current and future challenges of data growth, preservation, and management, and in supporting NSF’s new data management plan requirements. CISL data-centric CI design logically positions computer, disk, and tape-based storage systems to provide an efficient, safe, reliable environment for hosting and serving datasets.
As noted, one of CISL’s main integration points with XSEDE is via Globus data services and Globus data services are used by CISL’s ESGF nodes for bulk data access and replication support. CISL recently deployed Globus Plus, a robust and user-friendly service that eases data sharing, allowing scientists to be more productive by spending less time on the minutiae of data transfers. Globus Plus services also allow data providers to easily deposit large volume data collections at NCAR for publication and access via ESG and ESGF systems. The Globus Plus service at NCAR currently has a 1.5-petabyte capacity.

We believe that the national CI community would do well to focus more on the federation of data and data services across the nation’s CI rather than adopting the traditional HPC center mentality focused on primarily on big iron. The federation of cloud-based, data-centric computing coupled with scalable analytics forms a key component of NCAR/UCAR’s big data strategic vision, and we have begun exploring the migration of some of our traditional services to the cloud. These directions, in our view, are not only transformative, but probably inevitable given the broad societally and technologically driven trends around the creation and use of digital data.

**Future concerns**

As a discipline-specific CI organization, it has been CISL’s takeaway from the TeraGrid and XSEDE experience that, when it comes to federating CI resources, one size does not fit all. We support the notion that offering different levels of integration into a federated CI system is the most inclusive, productive and beneficial way forward. It may not be the easiest, however. CISL will continue evaluate new services and integration standards as these opportunities arise within XSEDE and other CI federations. In addition to ensuring that such services allow CISL to retain its domain-specific mission and identity within the broader federation, we are principally concerned about the scalability of large-scale CI federation in three areas: reliability, usability, and security.

**Reliability.** No utility, including grids or clouds, are immune to outages, and the natural tendency is for reliability to decrease as the size and complexity of a system increases. In 2013, widely reported outages at Amazon, Google Drive and Dropbox disrupted user access from hours to days, costing business users millions of dollars in revenue. Businesses or government entities dependent on real-time weather data will be highly susceptible to any cloud availability or QoS issues, even short term ones. Because the importance and time sensitivity of some scientific data, we would like to see technologies for 1) selectively replicating time sensitive or mission critical data across multiple platforms, and 2) auditing federated systems for vulnerabilities such as single points of failure and fault cascades.

**Usability.** One of the key barriers to data use (in particular "Big Data" use) is the usefulness and efficiency of current discovery and access systems. This issue broadly affects all of scientific CI. In particular, the process of evaluating data products for use by data consumers is commonly the most time consuming part of the scientific data discovery to analysis workflow.
This area should be noted as a concern for any future, federated CI development, not only for data resources, but for all types of federated resources.

**Security.** The integrity of personal information and scientific data is a major issue that is only complicated by the federation of CI. It is nearly impossible to guarantee 100% security and privacy protection against all possible threats. Still, scientific CI should not be complacent in the face of evolving threats as it designs future infrastructure: rather cloud-based and other federation technologies must be made much more secure proactively.