Cyberinfrastructure Plan
Colorado State University
December 1, 2014

I. Background

Colorado State University is a land-grant university with particular strengths in the applied sciences, engineering, life sciences, agriculture, and natural resources. Traditionally, these strengths have required the most advanced cyberinfrastructure. Indeed, CSU was the first university in the U.S. to purchase a supercomputer, a CDC Cyber 205, in December 1981. CSU thereafter served as an NSF Phase I Supercomputer Center, served as the NSF awardee for the Westnet regional network, was a founding member of the Front Range GigaPoP, was the first institution in the region to connect to the Internet2 network, and has organized itself strategically and operationally in the area of cyberinfrastructure to support research and education.

Many of our research areas involve 'Big Data,' some so voluminous they cannot be transported across the most advanced networks, thus requiring local distillation or even sampling of data in subsets. As examples, much of our research involves high-resolution, multi-wavelength satellite imagery as well as detailed, high-throughput 'omics data, including genome sequencing, proteome and metabolome profiling, and large-scale phenomic data. Our HPC system is capable of generating more than a terabyte per second of data. There is an urgent need for networking to 'catch up' to these needs, as it now is the most significant impediment to big science using big data.

II. Cyberinfrastructure Framework

Figure 1 depicts a conceptual framework for cyberinfrastructure, beginning with the generation of data (e.g. HPC systems and other instrumentation). Subsequent to generation, data must be stored, sometimes locally, and transported over networks. Analysis via scientific visualization or other post processing is then typical. It is generally after this that new knowledge is generated and scholarly publications result. Then, the publications, data sets, and metadata are archived in a digital repository for discovery and access by others, and preserved for posterity. Interwoven among all of these elements is the middleware needed for authentication and authorization to access these systems and data. All of this cyberinfrastructure is critical to enable the reuse of data to speed scientific discovery.

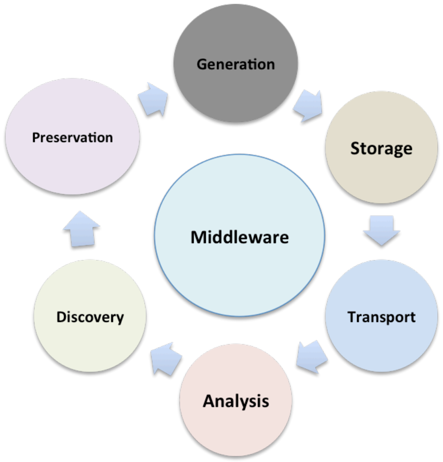

Figure 1 Cyberinfrastructure for 'Big Data'

III. Current Status of Cyberinfrastructure at CSU

A. Organizational Structure

In 2008, to reap the benefits of synergies between IT and information access, the VP for IT was jointly appointed as Dean of Libraries. Then, in 2011, to support Big Data, CSU merged Academic Computing and Networking Services (ACNS) with CSU Libraries. Librarians are consummate experts at dealing with *information*; while central IT staff are experts in *technology*. The synergies are illustrated in Figure 2. In addition, the faculty and research community are involved through the Information Science and Technology Center (ISTeC), which supports both education and research. ISTeC, a central organization representing all colleges on campus, reports to the VP for IT through CSU Libraries.

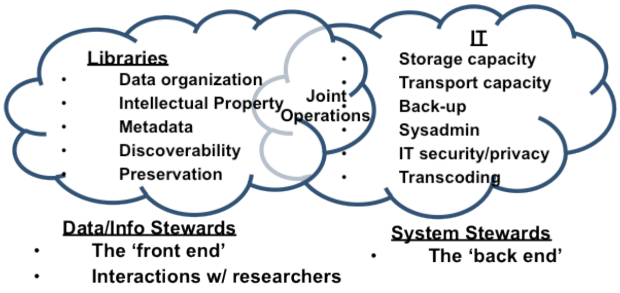

Figure 2 Information Infrastructure

B. Networking Infrastructure

i. Wide Area Network (WAN) – CSU obtains Internet access from the Front Range GigaPoP (FRGP) in Denver, Colorado, a collaboration that began in 1998. Access includes redundant links to two commodity providers: TeliaSonera and Level3; and various research networks: Internet2, CalREN, ESnet, and the Western Regional Network. CSU and its partners utilize the BiSON (Bi-State Optical Network) WAN, deployed using DWDM in a bidirectional self-healing ring, for transport to the FRGP. See Figure 3.

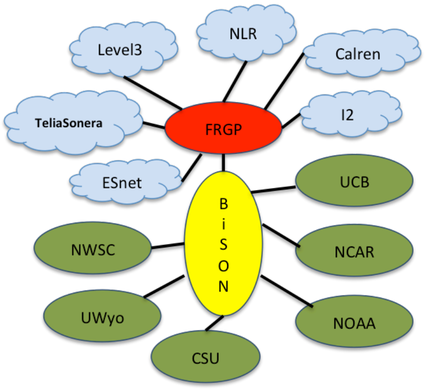

Figure 3 BiSON WAN

ii. Local Area Network (LAN) – CSU operates a fully redundant, self-healing 10 Gbps core backbone network with over fifty connections to campus buildings, some at 10 Gbps, most at 2 Gbps, and a few to smaller buildings at 100 Mbps (all are being upgraded to 2 Gbps), as shown in Figure 4. These connections serve traditional needs, for generic traffic. All major buildings are well covered with Wi-Fi.

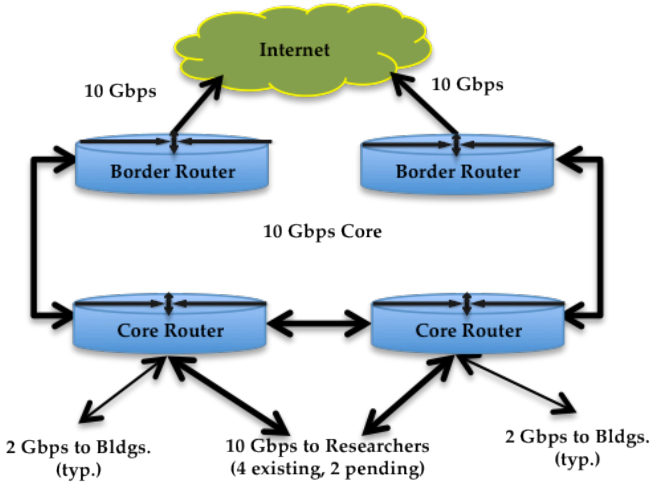

Fig. 4 LAN - 10 Gig Core

iii. Software Defined Networking (SDN) – CSU implemented DYNES via a grant from Internet2, but the networking community has deprecated DYNES and is seeking alternatives. As viable and practicable alternatives merge, we will deploy and integrate them into our infrastructure to support research and discovery.

iv. Research Connections – Four years ago, CSU began to establish dedicated 10 Gbps connections to research groups, based upon their need for ultrahigh speed. Currently, four have been deployed, two are pending, and we anticipate at least a doubling within two years. All of these use the shared research DMZ that also serves generic needs, a manifest bottleneck for performance.

v. Metrics and Statistics – CSU was one of the first institutions to participate in Internet2's measurement and performance activities, and remains current today with the most recent version of perfSONAR. Also, CSU's networking staff diligently alarms, monitors, and publishes comprehensive network traffic graphs for every backbone connection.

vi. Large-scale Storage Appliance (LSA) – CSU developed its own LSA solution in support of Big Data (see Figure 5). These appliances provide reliable, robust, fast, and affordable storage and preservation. The LSA is about one-fifth the cost of commercially available devices ($14k for the first 72 TB and $14k for each additional 96 TB of raw storage) and performs very well. This is a key element of our strategy to support storage and preservation of huge data sets, many of which cannot be moved across the fastest available networks. Currently, we have 17 such boxes deployed variously in our central IT environment.

Figure 5 LSA

vii. IPv6 – CSU has been experimenting with IPv6 since I2 made a block of IPv6 addresses available in 2006. As a result, we have a well-developed and mature knowledge base in the area of IPv6, and are ready to deploy it fully and to implement a transition plan when required.

C. HPC Infrastructure

CSU received an NSF MRI grant in 2009 to procure, install, operate, and manage an HPC system. A 12 TFLOP (peak) 1,248 core Cray XT6m system was installed on December 20, 2010. Formal academic classes and workshops in its use are continually being conducted. Over two hundred and fifty users are making very effective use of the system. This system is now congested, but plans are underway to add an additional

776 cores. The system is managed and operated by central IT, is under maintenance and support from Cray, and is performing extraordinarily well.

However, much more high-performance computing is done off campus than on campus. For example, CSU is the largest user of NCAR's HPC systems, and CSU researchers heavily utilize other supercomputer centers, including the OLCF and the NERSC. Greater on-campus and off-campus network capacity to and from the Cray, and to and from external resources, is certainly needed.

D. Digital Repository

Due to its extensive experience operating a digital repository, CSU hosts the Shared Services Digital Repository, encompassing nine campuses in Colorado (AMC, CSM, CSU, CSUP, MSU, UCB, UCB Law Library, UCCS, and UCD). That activity is organized specifically to share expertise, realize economies of scale, and take advantage of geographically distributed data centers for storage and preservation. It is this platform and the coalesced organizational structure that we use to support data management, data curation, and preservation, particularly in response to NSF's requirement.

E. Middleware

CSU was an early participant in InCommon, and was the first institution in the region to deploy Shibboleth, as both an Identity Provider and a Service Provider. CSU has widely extended its use of Shibboleth over the past two years, to both external and internal entities, and was the first in the world to deploy shib for Electronic Books Library (EBL). We are considering participating in Internet2's TIER activity that is emerging nationwide to provide new and more robust IAM services.

F. Affine Groups

CSU was a founding member of the Front Range Consortium for Research Computing (FRCRC), encompassing four research universities (CSM, CSU, UCB, and UWyo), and three national labs (NCAR, NOAA, and NREL). This activity was so successful it was expanded to the additional states of Utah and Idaho last year, and constituted as the Rocky Mountain Advanced Computing Consortium (RMACC). This activity now supports HPC events and symposia, and sharing information, expertise, networking, and computing cycles. See http://rmacc.org for more information about this collaboration that encompasses advanced networking as well as HPC.

IV. Plans for the Future

Networking Upgrades – The most severe need is to continue to progress with the networking upgrade funded in part by our NSF CC-NIE grant. That infrastructure should achieve fruition fully in calendar year 2015. Henceforward, we would then continue to upgrade links to and within buildings on our commodity LAN, and continue to add devices that require ultrahigh capacity into the Research DMZ.

Security Architecture – CSU has implemented a "Research DMZ" by deploying a dedicated infrastructure, located at the campus network perimeter, specifically engineered to accommodate ultrahigh bandwidth requirements of our research community. The Research DMZ is directly attached to a Juniper MX480 router, capable of providing judiciously tailored port filtering to protect high-end research systems with minimal performance impact. We believe this is a prudent balance between ensuring the highest end-to-end throughput while reasonably mitigating the risk of compromised systems.

This scalable architecture also enables CSU researchers to access advanced WAN services as they become available, such as 40 and 100 Gigabit infrastructure, dynamic virtual circuits and software defined networking (SDN) environments. It also supports higher fidelity performance monitoring by isolating research traffic from the routine data transfer needs of the academic and business side of the campus.

Wireless Infrastructure – In addition to providing the Research DMZ, the core campus data network is also being upgraded to better accommodate the insatiable appetite for mobile accessibility. Coupled with the continuation of enhancing both the speed and redundancy of connections to campus buildings and increasing the speed and density of wireless access points, thanks to a new, substantial recurring investment in capital and staffing, the campus is now well-positioned to meet the demands of our mobile community.

'Green' aspects – The campus backbone network was re-architected to move from a collection of five core routers to only two. This has obvious implications regarding the need for power and cooling in secure, environmentally controlled facilities. The enhancements described in this proposal fit nicely into that same reduced footprint, consistent with the university's "green" philosophy. Moreover, these devices are housed in CSU's 'green' data center, a well-managed, energy efficient data center. Finally, rotating storage UPS units are used, to obviate the need for use and disposal of heavy metals in battery-powered UPS units.

C. Storage

Unprecedented capacity in storage systems will be needed. Fortunately, we have an excellent device, our LSA, to meet these needs. We have added such storage to our Cray HPC, our digital repository, and to fifteen additional systems. Our plan for general storage for research data is two-fold: 1) we have developed a central virtualized server/storage/backup infrastructure that offers storage capacity at 30¢/GByte/year (not backed up) or 55¢/GByte/year (backed up), and 2) we have offered and will continue to offer workshops to campus IT support staff to build their own LSAs. In the summer of 2012, we offered this workshop ('Build your own Large-scale Storage Appliance') to seven institutions of higher education in Colorado, and it was very successful. Moreover, several LSAs have been deployed locally on campus too.

D. HPC

CSU installed an NSF MRI-funded Cray on December 10, 2010. It was expanded to a full cabinet two years ago, and now supports 2,016 cores. However, it has now reached its capacity limit, and we are in the process of seeking internal funding to upgrade it to its next generation that would approximately double its capacity, needed immediately, and projected to meet demand for about a year. ISTeC has been approved to submit another NSF MRI proposal for a new HPC system due January 2015, this time supporting accelerators including GP-GPUs and MIC. This system would be deployed circa December 2015, and support HPC needs for about another five years. Discussions are also underway to offer the service to researchers on campus under the 'condominium' model, as shared infrastructure, centrally supported and managed, for sustainability. Moreover, there is additional capacity off campus available and coming on line (e.g. Blue Waters, NERSC, OLCF, etc.). To meet this significant increase in demand, it is likely that a multi-faceted approach, involving progress in all of these areas, will be needed.

E. Shared Digital Repository

We recently adding about 50 TBytes of storage to our shared digital repository, and this has provided 'breathing room' to allow us, together with our partners, to plan for the next phase of a digital repository. Burns is chairing a committee in the Colorado Alliance of Research Libraries to identify a new, larger shared service digital repository supporting more than the nine current institutions in our extant shared digital repository. The two solutions being explored are Fedora/Islandora or Hydra or Dspace.

F. Middleware

We are in an excellent position regarding Middleware, having joined InCommon, and deployed a single shib Identity Provider, and numerous shib Service Providers on campus. We are also gainfully using Grouper, and are considering participating in Internet2's TIER activity.

Glossary of Acronyms

ACNS – Academic Computing and Networking Services

AMC – Anschutz Medical Center

ARL – Association of Research Libraries

BiSON – Bi-State Optical Network

CIFER – Common Identity Framework for Education and Research

CSM – Colorado School of Mines, Golden Colorado

CSU – Colorado State University

FRCRC – Front Range Consortium for Research Computing (members are CSM, CSU, NOAA, NCAR, NREL, UCB, and UWyo)

FRGP – Front Range GigaPoP

GP-GPU – General Computing on GPU (Graphics Processing Unit)

HPC – High Performance Computing

I2 – Internet2

IAM – Identity and Access Management

ISTeC – Information Science and Technology Center
KIM – Kuali Identity Management
LAN – Local Area Network
MIC – Intel's Many Integrated Core architecture
LSA – Large-scale Storage Appliance
LOCKSS – Lots of Copies keep Stuff Safe (a digital preservation framework)
MSU – Mesa State University, Grand Junction Colorado
NCAR – National Center for Atmospheric Research
NERSC – National Energy Research Scientific Computing Center
NLR – National Lambda Rail
NOAA – National Oceanic and Atmospheric Administration
NREL – National Renewable Energy Laboratory
NWSC – NCAR/Wyoming Supercomputer Center, Cheyenne, Wyoming
RMACC – Rocky Mountain Advanced Computing Consortium
STEM – Science, Technology, Engineering and Mathematics
TIER – Trust and Identity in Education and Research, an Internet2 initiative
UCB, UCCS, and UCD – University of Colorado at Boulder, Colorado Springs, and Denver
UWyo – University of Wyoming, Laramie, Wyoming
WAN – Wide Area Network