Introduction

The campus cyberinfrastructure (CI) is the connective tissue of technologies, data, applications, and people that enable collaborative research, teaching, and learning. The technologies include inter-networking within and among the University's campuses, national and international research partners; high performance computing; and the data collected by instruments, created through simulation, and shared by and with research partners. This document updates a 2015 CI plan[1] and summarizes the strategy, principles, and plans that have been developed and continue to evolve in support of developing this CI.

Stakeholder Engagement and Governance

Key research stakeholder engagement is formalized in the Research Computing Executive Committee (RCEC), consisting of senior research leadership and school deans; the faculty-led Shared Research Computing Policy Advisory Committee (SRCPAC) which reports to it; and the Information Technology Leadership Council, an advisory group to the Chief Information Officer. Further governance mechanisms have established a number of key policies to ensure information and network security, efficiency, capacity and reliability through a University-wide administrative policy initiative.[2]

As a founding member of NYSERNet in 1985 and with continuous representation on its board of directors, Columbia has been deeply involved in the creation and governance of research and education community shared services. NYSERNet has successfully developed several updates to our regional network connectivity to national and international partners, a robust New York City-wide dark fiber network connecting Columbia to the nation's preeminent international R&E colocation facility, and a statewide dark fiber network infrastructure and shared data center.[3]

General Approach to CI Development

Our approach to developing the campus CI is building networks and computational services to *support* research, not *as* research. Columbia has many faculty that are on the cutting edge of many parts of CI research, but our goals for campus CI are: to provide dependable resources that serve all faculty and students; to minimize friction for research data flows; and to provide reliable, convenient and accessible network and computing infrastructure using proven technologies. Our approach is to be fast-followers rather than pioneers, watching to see when new technologies and approaches mature sufficiently to provide value and reliability. This approach includes looking to cloud services — both consortial and public — as potential alternatives to on-campus resources, something that is especially important in our congested, urban environment where both physical space and energy are not cost-effective.

High Performance Computing

Current State. We've run a centrally-managed shared HPC service since 2009, governed by a faculty operations subcommittee of the aforementioned SRCPAC. Our initial shared cluster started with roughly three TFLOPS, initially serving astronomers and statisticians, managed by professional IT staff and

[1] "Columbia University Campus Cyberinfrastructure Plan 2015." Columbia University Academic Commons [distributor], March 2015, https://doi.org/10.7916/D8J38RF0
[2] https://policylibrary.columbia.edu/
[3] https://www.nysernet.org/

designed and measured to meet green targets for energy efficiency.[4] This service has continued to grow in scale and scope, with University, Federal[5] and New York State[6] support. We have since implemented three additional HPC clusters, increasing capacity to 256 TFLOPS, in addition to significant GPU capacity.

The faculty subcommittee has directed the creation of policy for researchers based on their buy-in commitments, a free use tier to allow experimentation with HPC services by faculty and students and use of the HPC service for instruction in graduate and undergraduate courses in scientific computing.

In addition to the on-campus shared HPC service researchers may have access to national resources such as XSEDE.[7]

Plan. The impact of big data being felt across the disciplines is significant. The need is obvious for CI to enable major initiatives such as the Data Science Institute,[8] the 2016 opening of the 450,000 sq. ft. Jerome L. Greene Science Center which houses the Mortimer B. Zuckerman Mind Brain Behavior Institute[9] and the Presidential Initiative in Precision Medicine,[10] to name a few.

CI-dependent innovation has emerged across all disciplines with expanded research requirements and new facilities on the five University campuses (Morningside, Medical Center, Manhattanville, Lamont-Doherty Earth Observatory, Nevis Labs); all have high performance computing and data analysis needs. SRCPAC works on developing shared approaches to meet these needs, including exploring use of the public cloud, understanding when and how to move computational workload to national and consortial facilities, and investigating approaches to energy efficient computing that will extend the life of our current data center infrastructure. In 2018, the University funded data center high-density cooling improvements, ensuring capacity for the next several generations of campus HPC.

Data Storage and Archiving

Current State. A 2008 e-Science Task Force[11] identified three development goals: to prepare for strengthened NSF and NIH data-sharing mandates, to expand the University Libraries' Academic Commons service to collect and preserve faculty scholarship and improve accessibility through search tools, and to develop a long-term plan to fully support federal agency requirements for data preservation and access. This service and strategy development[12] has led to continued growth of the Academic

[4] Work supported in part by the New York State Energy Research and Development Authority (ST-11145). "Columbia University Advanced Concepts Data Center Pilot: Final Report." Columbia University Academic Commons [distributor], March 2015, https://doi.org/10.7916/D87P8X8W

[5] NIH Research Facility Improvement Grant 1G20RR030893-01

[6] New York State Empire State Development, Division of Science Technology and Innovation (NYSTAR) contract C090171

[7] CUIT Research Computing Services, https://cuit.columbia.edu/shared-research-computing-facility#/text-9854, https://www.xsede.org/

[8] https://datascience.columbia.edu

[9] https://zuckermaninstitute.columbia.edu/

[10] https://precisionmedicine.columbia.edu/

[11] Neal, James G. and Renfro, Patricia E. "eScience Task Force Final Report." Columbia University Academic Commons [distributor], February 2009. http://dx.doi.org/10.7916/D8765C89

[12] Bose, Rajendra K. and Nurnberger, Amy L. "Columbia's Evolving Research Data Storage Strategy." Workshop on Research Data Management. Arlington. March 2013. https://doi.org/10.7916/D8BV7R3V

Commons research repository. To support this growth, the University Libraries has created the Digital Preservation Storage System to provide long-term storage and access to digital library collections and support additional repository services. The Libraries also provide training, consultation, and support for researchers' development and implementation of data management plans. Data classification and security policies have been updated and promulgated to the campus community in order to meet regulatory and information security requirements.

There are several data storage options available to researchers, depending on their needs.[13] Many researchers continue to rely on low-cost local storage solutions in the absence of any departmental, school or University-provided alternatives of comparable cost and utility. While researchers are often aware of the risks inherent without offsite backup in place, available solutions for this are often limited and/or perceived as too expensive.

Plan. As data grow, we continue to develop affordable ways to protect and archive raw data for potential future re-analysis, means of sharing massive datasets over networks, search and discovery, and increasing concerns around data security and privacy. We continue to develop capabilities in this area, especially with respect to consortial[14] and public cloud services in order to support both short-term and long-term preservation, access and reuse of research data.

Network

Current State. The leading principle of Columbia's network architecture is to be *frictionless*. That is, gaining access should be immediate and trivially easy, moving data across the network should not be limited and the end-to-end communication principle of the Internet Protocol is maintained.[15] As such, Columbia continues to maintain a relatively open and unimpaired "free love" network well into the 21st century.[16] User end-devices are allocated publicly-routed IP addresses. There is no campus border firewall and thus no need for a Science DMZ; services are walled off at a more granular microdomain level within the network. The fundamental design assumption is that the campus network, which connects academic, administrative and residential users (including faculty and graduate students in over 100 University-owned neighborhood apartment buildings) *is* the Internet. This is not to say that we have no network security. In fact, automated netflow[17] and system access log analysis[18] is continuously performed, including automated removal of compromised hosts from the network. A bandwidth quota system protects against excessive use of commodity Internet capacity[19] with no limits placed on Research & Education network usage.

---

[13] https://research.columbia.edu/content/research-data-storage
[14] CU contributes in many consortia including: Digital Preservation Network, APTrust, HathiTrust, DuraSpace, etc.
[15] Saltzer, Jerome H., David P. Reed, and David D. Clark. "End-to-end arguments in system design." *ACM Transactions on Computer Systems (TOCS)* 2.4 (1984): 277-288. https://doi.org/10.1145/357401.357402
[16] Kundakci, Vace. "Free Love and Secured Services." Educause Review (2002): 66-67. Retrieved April 11, 2019 from https://www.educause.edu/ir/library/pdf/erm0266.pdf
[17] Rosenblatt, Joel. "PAIRS/Bayesian IDS: Finding Bad Actors without Looking at Content." Educause Security Professionals Conference. 2011. http://www.educause.edu/sites/default/files/library/presentations/SEC11/SESS10/PAIRS%2BSPC%2B2011.pdf
[18] Selsky, Matt, and Daniel Medina. "GULP: A Unified Logging Architecture for Authentication Data." *LISA*. 2005. http://static.usenix.org/legacy/events/lisa05/tech/full_papers/selsky/selsky_html/
[19] http://policylibrary.columbia.edu/network-protection-policy

With network freedom comes network responsibility: anti-spoofing (BCP 38) is performed throughout the campus as is botnet membership detection and mitigation, so as not to have our network used as a DDoS source.

Campus IP network services are centrally managed all the way to the end-user's network jack (or WiFi access point) in the majority of buildings, enabling end-to-end visibility for performance management and security. All network devices are SNMP-monitored. The standard for new construction and renovation[20] specifies ubiquitous high-density WiFi (802.11ac) and Category 6A jacks in research lab spaces. The campus distribution network, which extends to the R&E and commodity Internet colocation facilities, currently operates at 40 Gbps, with building switches uplinked at 1 or 10 Gbps. Some older construction is still saddled with Category 3 wiring and has limited capacity, although the focus has been on investing in high-capacity WiFi in those areas.

The NYC dark fiber metro area network provides robust 40 Gbps connectivity between the Morningside, Manhattanville and Medical Center campuses and wide area dark fiber circuits implement a 10 Gbps ring connecting the Nevis Labs (Irvington, NY) and Lamont-Doherty Earth Observatory (Palisades, NY) campuses to our NYC colocation facilities and Morningside campus.

Dual-stack IPv4/IPv6 peerings are in place with both commodity ISPs (Cogent, GTT) and the R&E networks (via NYSERNet). IPv6 routing is currently available to subnets in the computer science department, based on researcher interest.

Also available is a 10 Gbps Amazon Web Services "Direct Connect" private peering to support research in the AWS public cloud environment.

Customized wavelengths and dark fiber links are provided on a limited basis for truly special cases.[21]

Network availability and performance is continuously monitored, with incidents handled by our Network Operations Center and on-call staff. Iperf[22] is used to commission new installations and as a debugging tool when diagnosing end-to-end network performance concerns.

Plan. Planned wide-area network improvements include continuing to leverage NYSERNet's leadership in advancing connectivity with peer research and education networks[23] as well as commercial providers via various peering arrangements with, for example, DE-CIX, Equinix, and NYIIX.[24]

We continue to address networking in legacy Category 3 buildings primarily with WiFi expansion and new wiring as research labs are renovated.

For the time being, our dark fiber network investments and use of IP networks seem sufficient, but we remain ready to support researchers who might need more specialized connectivity via wavelengths and the like, although we maintain a healthy level of skepticism regarding the actual (vs. perceived) need for

[20] https://cuit.columbia.edu/sites/default/files/content/CUIT%20Network%20Infrastructure%20Technical%20Design%20Requirements.pdf
[21] See, for example, NSF award CNS-1827923. https://cosmos-lab.org/
[22] https://github.com/esnet/iperf
[23] https://www.nysernet.org/r-and-e/
[24] https://www.peeringdb.com/net/12186

these services that complicate the end-to-end model. We are well-positioned to take advantage of these services given our presence in the NYSERNet colocation facility in Manhattan.

Supporting Shared Services

Current State. In order to support collaboration among Columbia researchers and their national and international peers, we participate in the InCommon Federation, having SIRTFI certification,[25] and implement the usual identity attributes shared across our community. All students, faculty and staff are automatically on- and off-boarded in our identity and access management service based on our systems of record. Research colleagues, contractors and others may also be granted access through a Delegated Identity Administration capability. To enhance identity assurance, we also employ multi-factor authentication (MFA). Collaboration services such as GSuite for EDU are available for use by most researchers, with the exception of those subject to HIPAA and ITAR regulations.

Central IT's Research Computing Services team[26] in conjunction with the University Libraries[27] and the Office of the Executive Vice President for Research provide centralized services[28] that include:

- Secure Data Enclave (SDE) for restricted use datasets;
- Cloud consulting services;
- Research data management consulting;
- Enterprise license for Globus for research data transfer;
- An electronic lab notebook cloud service;
- Piloting a platform for code reproducibility and archival (begun in 2019); and
- The *Foundations for Research Computing*, described below.

Education

Current State. Researchers historically learned computational methods on-the-job and through short HPC and data management workshops and special interest groups coordinated by CI staff from IT and the Libraries, and through colloquia and formal coursework offered by faculty in several disciplines.

In May 2018, the RCEC supported creation of the *Foundations for Research Computing*[29] program to address the need to provide informal training for Columbia University graduate students to develop computational skills. Beyond training, the *Foundations* program aims to create a computational community at Columbia, bringing disparate researchers together with the common thread of computation. Part of the Foundations for Research Computing is the membership in and adoption of *The Carpentries* to build "global capacity in essential data and computational skills for conducting efficient, open, and reproducible research."[30]

Plan. With the oversight of the RCEC, we plan to expand the current offerings of the Foundations for Research Computing to provide more intermediate and advanced trainings and discipline-specific curricula over the next 2-3 years.

---

[25] https://incommon.org/custom/federation/info/all-idps-certified.html
[26] https://cuit.columbia.edu/cuit/research-services
[27] https://library.columbia.edu/services/research-data-services.html
[28] https://data.research.columbia.edu
[29] https://rcfoundations.research.columbia.edu
[30] https://carpentries.org