Identifying Research Collaboration Challenges for the Development of a Federated Infrastructure Response

Maureen Dougherty
maureen.dougherty@ernrp.org
Eastern Regional Network
USA

Michael Zink
zink@ecs.umass.edu
University of Massachusetts, Amherst
USA

James Barr von Oehsen
jbv9@rutgers.edu
Rutgers University
USA

ABSTRACT

In this paper we present the key collaboration challenges and recommendations identified by targeted research communities during the Eastern Regional Network (ERN) [7] Architecture and Federation Virtual Workshop[6], for validation of the base design of the ERN Federated OpenCI Labs collaborative infrastructure model. The workshop was designed to stimulate open discussions surrounding key aspects of collaborative scientific research and workflows. A brief summary of the key data gathered is provided here. The findings from this workshop have led to a re-evaluation of the design of ERN Federated OpenCI Labs infrastructure.

CCS CONCEPTS

• **Theory of computation → Interactive computation**; • **General and reference** → Validation; • **Social and professional topics → Computing / technology policy**; • **Networks → Network design principles**; • **Security and privacy → Security protocols**.

KEYWORDS

Research Computing, Federation, Cloud Services, Edge Computing, Core Facilities

ACM Reference Format:
Maureen Dougherty, Michael Zink, and James Barr von Oehsen. 2021. Identifying Research Collaboration Challenges for the Development of a Federated Infrastructure Response. In *Practice and Experience in Advanced Research Computing (PEARC '21), July 18–22, 2021, Boston, MA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3437359.3465594

1 INTRODUCTION

The Eastern Regional Network (ERN)[7] was formed with the vision to simplify multi-campus collaborations and partnerships in the Northeast United States, in order to advance the frontiers of research, pedagogy, and innovation. This vision led to the design of a federated collaboratory that requires the development of many layers of abstractions ranging from hardware, networking, federation architecture, scientific workflows, and domain-specific models and tools to enable collaborative discovery. The ERN Architecture and Federation working group (AFWG) explores what the "federated collaboratory" abstraction layers might look like from both a hardware and software perspective as well as what federation should look like as we strive for a seamless collaborative sharing experience. Based on collective data from the ERN Steering Committee, six ERN working groups, and partnering research and networking collaborators, an initial federated infrastructure, the ERN Federated OpenCI Labs (OCL), a private cloud architecture, was developed. OCL was designed around the Computer Science Materials Discovery, and Structural Biology science drivers. In addition, OCL's goals are the support of under-represented/under-resourced institutions (Broadening the Reach) and the optimal use of existing resources including New England Research Cloud (NERC) [2] and Open Infrastructure Lab (OIL) [4] as well as NSF funded projects like FABRIC [3], Open Storage Network (OSN) [8] and Massachusetts Open Cloud (MOC) [5]. Critical to the model's success is the assurance that its foundation is solid and that its components properly address the collaboratory needs of the targeted research communities, particularly regarding remote access to scientific instruments and resources. In December 2020, the ERN AFWG held an information-gathering workshop to obtain additional insights from the ERN stakeholders, and any new key factors were incorporated into the OCL design. This report provides a summary of the primary challenges and findings identified by this workshop.

1.1 Open Cyberinfrastructure Collaboratory (OpenCI Labs)

Merging what we learned from the working group meetings, workshops, and the community, our goal for the OpenCI Labs project has evolved into building an instrument that interconnects the research instruments on our campuses through a federated private cloud platform, a secure but open gateway to our campus CI ecosystems, coupled with distributed specialized data transfer nodes that provide near data computing and advanced networking capabilities placed within close proximity to a research instrument (edge cloud services). This ERN effort driven by the needs of the Structural Biology, Materials Discovery, and Computer Science communities brings together people from all the major stakeholder groups and partner sites across the region and beyond to focus on standing up these services in order to revolutionize the way the research and education communities across the nation collaborate, access research instruments, and share data, ultimately leading to new and exciting research and education endeavors.

CI Building Blocks, as shown in Figure 1, are the hardware and software components of OpenCI Labs. Each building block is designed to be a standalone system or can be combined with other building blocks depending on the uses cases and campus needs.

PEARC '21, July 18–22, 2021, Boston, MA, USA
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8292-2/21/07...$15.00
https://doi.org/10.1145/3437359.3465594

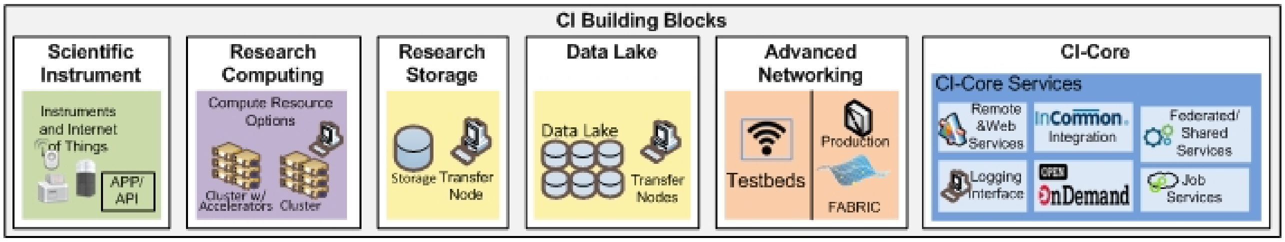

Figure 1: Overview of the ERN OpenCI Labs Building Blocks

The core services of OpenCI Labs is designed to allow for large scale distributed federation across multiple campuses, centers, and regional network providers. Each of these Building Blocks leverages OpenStack, making it easier for each of the partner sites to build an OpenCI Labs configuration. The software is designed to provide future proofing, ease of support and maintenance, procurement, and installation. Below we list each of the OpenCI Labs building blocks and a brief description of their purpose.

- Scientific Instrument - scientific or research hardware devices which can include any software needed to control the system (API or virtualized app). The instrument API and App are wrapped by common layers to insert into the site core and be managed as part of the federation. The goal is to connect and make the scientific instrument run 24x7, use remotely to help build collaborations, and provide access more efficiently.
- Research Computing - Research Compute resources (Login node, Scheduler, HPC/HTC, etc..). Integrates with the CI Core Job Services module to provide a consistent interface for all compute infrastructure including the ability to support multiple schedulers and data to job or job to data operations. Designed as open access for under-resourced colleges and universities.
- Research Storage - Includes a Transfer Node and Storage resources. They could include NVMe, in-Memory, Parallel, or future storage technologies. Need to be flexible to address future technologies.
- Data Lake - Transfer Nodes and Storage, generally large volumes and likely object storage today, but the integration will support future technologies as well. Any specific technology can be integrated as a data lake building block by developing a shim to integrate with the building block stack.
- Advanced Networking - Includes both production level and experimental (P4, FPGA based network switches or NICs) paired with PerfSONAR and logging interfaces to incorporate into the federation (standard configurations, etc). The SDN interface is wrapped to provide consistent controls as required by the Transfer Node or Job Services. The capability to connect to a FABRIC rack will also be incorporated. Potential for on-demand SDN.
- CI-Core - Provides CI Core Services that wraps one or more of the building blocks and enables it for Federation Integration. CI-Core Services includes the software stack common to all building blocks and acts as the federated service center for OpenCI Labs. Core service components include:
  – Open OnDemand (OOD) - Provides all the capabilities of OOD today, but adds federating access.
  – InCommon - Provides the federated authentication.
  – Remote and Web Services - Provides remote access for instruments and applications with web access leveraging federated integration. Allow for API integration and remote access to instrument control.
  – Logging Services - Integrated central logging facility that allows the building blocks to consolidate logging for security auditing for all layers (network, system, app).
  – Job Services - Provides a consistent interface for compute and job storage infrastructure with advanced capabilities around data to job and job to data transitions across the network leveraging workflow engines and secure network pathways.
  – Federated Shared Services - Are the overall services provided by the OpenCI Labs Federation.

2 ERN ARCHITECTURE AND FEDERATION WORKSHOP

2.1 Overview and Goals

Supported by the National Science Foundation CC* CRIA Eastern Regional Network planning grant (OAC-2018927), the ERN Architecture and Federated Virtual Workshop was held December 2-4, 2020 that included over 60 participants representing 37 different universities and organizations. The workshop's goals were to gather input from ERN stakeholders about their architectural requirements necessary for collaboration and scientific advancement, and to hear from architecture experts about approaches and solutions for the major building blocks of the architecture. To generate the most impact for the participants and the OCL design, the workshop was focused on the following fundamental categories identified by the AFWG: 1) Science Drivers 2) Authentication, Authorization and Accounting 3) Federated Infrastructure 4) Data – Storage, Sharing, Transfer, and 5) Network Connectivity. For each topic, experts and researchers in the areas of interest presented their efforts, sharing their insights and experiences. Breakout sessions following the conclusion of topic presentations provided additional opportunities for deep dive discussions. The format was designed to share information on collaborative efforts, best practices, resources, services and tools that could be beneficial to participants while gathering information that could lead to new options and adjustments to the ERN Federated OpenCI Labs design.

2.2 Sessions

The following subsections provide brief summaries of the workshop sessions.

2.2.1 Science Drivers. The Science Drivers sessions identified several areas of concern. Challenges that have implications for the other categories covered by the workshop will be mentioned in the respective sections. Most prominently, remote access to resources, including scientific instruments located in labs (edge), in a convenient and secure method was identified as a challenge. Remote resource access is complicated, complex and can be limiting to the point that it is not feasible. Not all scientific instruments located in labs, i.e. edge resources, have the capabilities to be remotely accessed. Those that can, may not have a manufacturer supplied interface, or if it does, it may not allow external network domain access. Local security and risk management policies may prevent access to resources (data, compute and instruments), by remote researchers, initiating access externally from the institution's network without specialized configurations. A local institution's compute environment, including software, may not be reproducible at remote location. Finally, some Broadening the Reach (BTR) institutions are unaware of available resources, and would need assistance to access and utilize local, regional and national resources. These were several access issues discussed.

2.2.2 Authentication, Authorization and Accounting. As noted in the science drivers session, authentication, authorization and access (AAA) policy and procedures implemented by the local institution can increase collaboration barriers. Stakeholders indicated that a federated infrastructure should limit the introduction of additional rule sets or policies to ensure local institutions are empowered to provision their local resources, and authorization and access remains delegated to the local resource and service provider/owner. Incorporating federated identity, service access and authorization should not require a separate guest account for each collaborating researcher. Existing NSF funded projects like InCommon, CILogin and Grouper, were deemed logical options for addressing this concern. Also discussed was the importance of the ability of AAA to traverse network domains providing seamless data migration, workflow operations and remote access/control to scientific, edge, instruments.

2.2.3 Federated Infrastructure. This area raised several discussion points. Of significant impact were workflows and the ability to reproduce analytical environments. The OCL model would include the implementation of scientific workflows for real time, remote provisioning, to support end-to-end scientific analysis that would include the edge device, the scientific instrument, computational resources, and the networks connecting these components. The workflows could launch containers or virtual machines, to provide a reproducible environment which can run on a pre-configured base set of compute requirements. Stakeholders indicated that the impact on analytical workflows regarding proximity of scientific instrumentation and computational resources should be part of design consideration. The federation should provide generalized containers and allow researchers to build and share containers designed for specific analytical software, methodology and reproducibility of analytics. Training and education in best practices and effective use would need to be available, particularly for the BTR institutions.

2.2.4 Data - Storage, Sharing, Transfer. Discussions throughout all sessions touched on various forms of data support and requirements necessary for collaboration and the creation of research workflows. The data requirements and challenges are diverse and complex, covering such topics as storage and archival needs, data migration barriers, standardizing data collections, data management for large data sets, data accessibility, data bases, and establishing policies and guidelines to address ownership and liability. Technology advances produce larger datasets requiring storage, migration, analysis, archiving, and curation. Sustainable, increased storage was a reoccurring theme. Discussions revolved around how best to address these services while ensuring data is secure, searchable, and accessible in a cost effective and sustainable manner. Of note was the need for a database for curated datasets based on defined standards and metadata providing searchable, accessible datasets for the Materials Discovery community, comparable to the resources currently available globally to the Structural Biology community through the Protein Data Bank [1].

2.2.5 Network Connectivity. Researchers indicated some of the key network connectivity issues that they have experienced involve insufficient bandwidth and latency issues, workflow path generation, and expertise in network optimization, efficiency and troubleshooting. Bandwidth, latency and network bottleneck issues can significantly impact data transfer rates and real-time remote instrument control and/or image processing, as to make them impractical. It was noted that USB drives for data transfers instead of software transfers across the network are still common at several institutions. Workflows involving remote resources may require traversing numerous network domains and institutions, necessitating the creation of an authorized, authenticated network path across those domains. Ideally this would be automatically configured at the launch of the workflow. Access to expertise and training to efficiently configure, optimize and troubleshoot the network, particularly for the BTR institutions, was identified as indispensable to address some of these limitations.

3 POLICY

A common, underlying element throughout all of the workshop sessions was policy. Concerns surrounding data sharing, ownership and risk management were raised during numerous discussions. Current institutions' security and risk management policies and guidelines present barriers making local access to resources challenging, and the concept of remote collaboration daunting. Examples of small, localized projects demonstrated the complexity encountered when collaborating using shared resources. For some, this included reviews of local and regional agreements with the creation of new use agreements between participating institutions. Ownership and AAA control of resources was also voiced. It was agreed that the local institution should maintain ownership and manage AAA services for their shared research resources. The ERN will need to establish policies and guidelines that address ownership and liability concerns surrounding data sharing and external federated data use, potentially creating data use agreement templates for ERN and community distribution. The ERN Policy Working Group (PWG) is tasked with the development of a comprehensive set of policies consistent with current university policies and requirements needed for a sustainable, secure, and compliant system. The initial concept is to generate a minimum set of policies that are clear, clean and as simple as possible, based on the principles of fairness, accountability, transparency and equity, while protecting the owner's rights at the university level. The PWG collaborates with all ERN working groups for input and insight. The ERN OCL includes these efforts and is designed to address future policy and guideline updates.

4 FINDINGS

The three-day workshop successfully identified numerous logistical and technical challenges facing researchers and their collaborative partnerships, particularly in areas of data, authentication, authorization, access, networks, workflows and online expertise/training. These will be evaluated for options and potential adjustment to the ERN OCL infrastructure design. Some of these issues may be addressed with collaborative partnerships with local, regional and national centers and NSF funded projects. Development of instrument interfaces enabling remote access to edge devices would require vendor partnerships more readily available to a federation than to an individual institution. ERN will be forming an Industry Advisory Board to aid in this endeavor and create new opportunities for all stakeholders. Strategic to the success of the ERN OCL will be our goals to reduce barriers, keep use and management simple, leveraging existing local policies and procedures, as well as utilizing existing applications and software for services instead of rewriting them. Development and deployment of a federated solution is to be pragmatic, and start with a small set of instruments to develop, test, analyze and understand the basics. By generating something prescriptive, we would have a richer knowledge that could be shared, replicated and integrated by the community.

5 USE CASE

After evaluating the findings and recommendations from the workshop and discussions with the ERN Science Drivers, a Proof of Concept (PoC) was initiated targeting remote scientific instrument access and utilization through the incorporation of a federated infrastructure testbed based on the ERN OpenCI Labs design. The execution of this PoC would validate several of the CI building block component designs, with adjustments made based on feedback from stakeholders, the Policy, Structural Biology, Materials Discovery and Architecture and Federation working groups.

This PoC is partnering with at least one select CryoEM lab and the scientific instrument industry vendor to achieve an automated computational analysis workflow leveraging NSF funded computational resources. Fundamental to the design is the access and control integration of the scientific instrument's interface. The industry manufacturer, ERN and researchers are in active discussions regarding interfaces and tools, existing and under development, which would provide secure remote access and control to their instrument. Workflows would evolve to establish data flows from the scientific instrument to targeted remote storage, and could include computational analysis of the datasets leading to live algorithm adjustments to the scientific instrument. Hurdles with authentication, authorization and access to the local, edge, scientific instrument by non-institutional researchers would be addressed through integration of the vendor API with software interfaces leveraging open source tools like InCommon and Open OnDemand.

Lessons learned and the ultimate results from all ERN PoCs will strengthen the ERN federated OpenCI Labs design, future proposals, and benefit communities beyond the structural biology community. This information will be shared with the research community through workshops, conferences, presentations and the annual ERN All Hands Meeting. The ERN encourages anyone interested in learning more about our collaboration efforts and how they can participate by visiting our website (https://ernrp.org) and sign up for our mailing list.

ACKNOWLEDGMENTS

The authors wish to thank all the members of the Eastern Regional Network Architecture and Federation Working Group for their contributions to the workshop and the ERN Steering Committee for guidance and support. This work was supported by the National Science Foundation CC* CRIA Eastern Regional Network planning grant (OAC-2018927).

REFERENCES

[1] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. 2000. The Protein Data Bank Nucleic Acids Research, 28:235-242. Retrieved May 19, 2021 from https://www.rcsb.org

[2] New England Research Cloud. 2021. New England Research Cloud. Retrieved April 21, 2021 from https://nerc.mghpcc.org

[3] FABRIC. 2021. The FABRIC Project. Retrieved April 12, 2021 from https://fabric-testbed.net

[4] OpenInfra Labs. 2021. OpenInfra Labs. Retrieved April 12, 2021 from https://openinfralabs.org

[5] MOC. 2021. The Massachusetts Open Cloud. Retrieved April 12, 2021 from https://massopen.cloud

[6] Eastern Regional Network. 2020. The Eastern Regional Network Architecture and Federated Virtual Workshop. Workshop Agenda. (December 2020).). Retrieved April 13, 2021 from https://drive.google.com/file/d/1KrFtC0Nb-z98bPb3sESv709lPNQ96Vxt/view

[7] Eastern Regional Network. 2021. The Eastern Regional Network. Retrieved April 12, 2021 from https://ernrp.org

[8] Open Storage Network. 2021. The Open Storage Network. Retrieved April 12, 2021 from https://www.openstoragenetwork.org