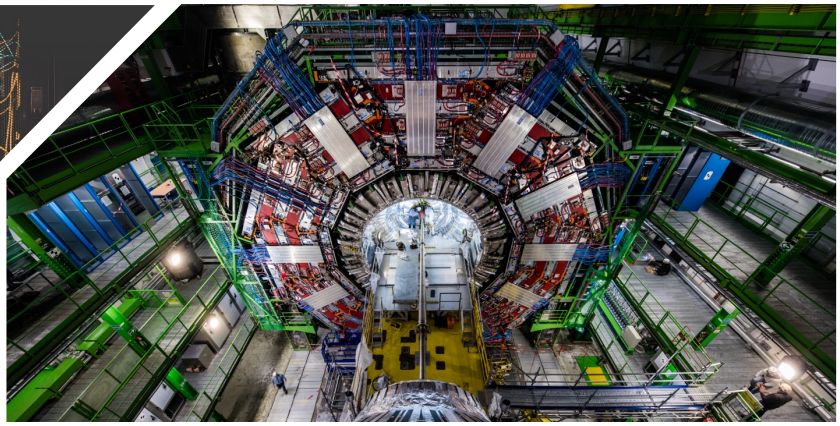
2019 NSF Workshop on Connecting Large Facilities and Cyberinfrastructure

Connecting Large Facilities, Connecting CI, Connecting People



http://facilitiesci.org

The workshop was funded by the National Science Foundation under grant #1933353

September 16-17, 2019
Alexandria, VA

Disclaimer. This material is based upon work supported by the National Science Foundation under grant #1933353. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the workshop participants and do not necessarily reflect the views of the National Science Foundation.

Cover. Designed by Rafael Ferreira da Silva (University of Southern California). Cover images credits: The Ice Cube Neutrino Observatory[1], The Large Hadron Collider[2], The Nathaniel B. Palmer at McMurdo Station (Kelly Cheek, National Science Foundation)[3], NEON Meteorological Flux Tower at terrestrial sites (Abraham Karam, NEON), Gray and Black Galaxy[4], Aerial Photography of Wide Green Grass Field (Stephan Müller)[5], Meeting photo (Facilities CI workshop)[6], and Cable Network (Taylor Vick)[7].

[1] https://www.nsf.gov/news/mmg/media/images/ICL_SmallDomBlue_Horizontal.jpg

[2] https://news.wisc.edu/content/uploads/2018/09/hadron-collider.jpg

[3] https://www.nsf.gov/news/mmg/media/images/NBP_H.jpg

[4] https://www.pexels.com/photo/sky-space-dark-galaxy-2150

[5] https://www.pexels.com/photo/aerial-photography-of-wide-green-grass-field-753869

[6] https://facilitiesci.github.io/2019/

[7] https://unsplash.com/photos/M5tzZtFCOfs

Report Authors:

Ewa Deelman (PI and Chair), University of Southern California
Ilya Baldin, RENCI, UNC - Chapel Hill
Brian Bockelman, Morgridge Institute
Adam Bolton, National Optical Astronomy Observatory
Patrick Brady, University of Wisconsin-Milwaukee
Tom Cheatham, University of Utah
Laura Christopherson, RENCI, UNC-Chapel Hill
Rafael Ferreira da Silva, USC Information Sciences Institute
Tom Gulbransen, Battelle, NEON
Kate Keahey, Argonne National Laboratory,
Marina Kogan, University of Utah
Anirban Mandal, RENCI, UNC-Chapel Hill
Angela Murillo, Indiana University-Purdue University Indianapolis
Jarek Nabrzyski, University of Notre Dame
Valerio Pascucci, University of Utah
Steve Petruzza, University of Utah
Mats Rynge, USC Information Sciences Institute
Susan Sons, Indiana University, CACR
Dan Stanzione, Texas Advanced Computing Center
Chaudhuri Surajit, Microsoft
Daryl Swensen, Oregon State University
Alexander Szalay, Johns Hopkins
Douglas Thain, University of Notre Dame
John Towns, NCSA
Charles Vardeman, University of Notre Dame
Jane Wyngaard, University of Notre Dame

Table of Content

Executive Summary

Summary of the Workshop

In 2015, the National Science Foundation (NSF) began to support a series of biennial workshops that bring together large facilities (LFs) and cyberinfrastructure (CI) projects to share common experiences and challenges, discuss potential collaborations, and identify opportunities for leveraging CI within the community. The 2019 NSF Workshop on Connecting Large Facilities and Cyberinfrastructure aimed to continue and advance this discussion, enable the exchange of CI solutions and challenges, and foster CI community building.

With a theme of "Connecting Large Facilities, Connecting CI, Connecting People," the 2019 workshop emphasized the need to facilitate collaborations among LF and CI projects and recognized the importance of the CI workforce to LF science missions. Specific goals of the workshop included: 1) identifying CI challenges that facilities face when supporting their science missions; 2) exploring the opportunities and obstacles to collaboration between LFs; 3) examining non-technical challenges that influence CI development; and 4) developing ideas for enhancing the CI workforce and building a community of CI professionals.

A steering committee of cyberinfrastructure experts from NSF LFs and selected CI projects defined the workshop agenda with input gathered through a pre-workshop survey. Close to 100 representatives from NSF LFs, major NSF-funded CI projects, selected cloud providers, and NSF staff participated in presentations, panels, and breakout sessions centered around the workshop goals.

Prior to the workshop, a survey was sent to the participants to gauge their willingness to share CI resource and expertise and to share workforce development efforts. The 32 survey responders expressed ambivalence with regard to sharing resources but enthusiasm at the prospect of sharing expertise. The most frequently cited obstacles to developing a shared services effort were limited funding and personnel as well as divergent missions and goals of individual LFs. When asked about their willingness to contribute to a shared workforce development effort, most respondents indicated that they did not have resources to put toward this type of activity. Obstacles to sharing in the area of workforce mirrored those to sharing resources.

In September 2019, to obtain a quick pulse of the community, workshop participants were asked to distribute a survey to their teams to gain information about CI knowledge sharing practices, incentives to following a CI career path, and feelings of belonging to the broader professional CI community. Seventy-eight respondents completed this survey. The survey revealed that the LF community relies on online sources for information seeking and knowledge sharing. Key findings in the area of job satisfaction were that the CI practitioners are motivated primarily by the "sense of purpose/helping science/the world" followed closely by the ability to conduct research and solve complex problems. Although many respondents felt a belonging to their particular community, LF, or project, the survey did not show broader community connections.

A community building effort was also launched by requesting workshop participants to provide "CI calling cards" with professional information, pictures, and responses to questions about recent CI frustrations and successes. Frustrations revolved around data management challenges, budgetary constraints, communication (e.g., changing mindsets), workforce (e.g., hiring and retention), technical aspects of LF work (e.g., planning and design, the tension between older existing and emerging technologies, and juggling heterogeneous CI). Successes included developing new systems or services, improving existing ones,

continuing to do good work throughout the year, and bringing people together via effective communication. These calling cards greatly facilitated group and one-on-one discussions. Workshop materials, including the presentations and calling cards, are available on the workshop website http://facilitiesci.org.

Summary of the Discussions

LF cyberinfrastructure challenges

LFs face a spectrum of CI challenges when supporting their science missions. They collect and disseminate very large, diverse, and heterogeneous data sets with different temporal and spatial scales. At the same time, CI and commodity technologies are rapidly evolving. LF mission focus and the need to mitigate long-term risk often mean that software is developed in-house rather than adopted from existing solutions or created through collaborations. Timelines and budgets also may drive LFs to pursue expedient solutions rather than examine potential new designs and technologies. The same pressures can limit LF willingness to seek and maintain collaborations that could increase CI reuse and adoption.

In addition to the technical challenges, LFs face ever-growing demands in the area of workforce training, development, and retention. They are constantly competing with industry to attract and retain CI talent and must often hire domain scientists, programmers, or administrators without the needed experience to navigate the complex CI landscape. Even as the CI workforce matures, quickly changing technology makes it difficult for individual CI practitioners and LFs to keep up with new capabilities.

Opportunities and obstacles to collaboration between LFs

LFs have limited resources to collaborate with each other and large CI projects as they must focus on delivering instruments, data, and software to their communities. Although collaborations are potentially desirable and beneficial, it takes time to discover appropriate partnerships, understand common interests, and maintain a collaboration until it starts bearing fruit. The nature and extent of CI collaboration may vary across the spectrums of CI developers' expertise in computer science and scientific domains. There are also often mismatches between LF construction timelines and CI projects. Finally, LFs frequently perceive risks with engaging in external collaborations during the construction phase, especially if the projects and programs do not have common NSF program officers.

As CI complexity increases, there is a growing need for sharing communication, coordination, information, and expertise across the facilities. The workshop identified simple sharing approaches that could benefit LFs in the short term. For example, sharing knowledge of existing LF architectures can inform new LF design, and CI project training opportunities could be shared with LF staff. Guidance on open source licenses, community data standards, workflows, and software development best practices could also help LFs develop interoperable CI components and more easily discover potentially useful CI services. As collaborations and trust amongst LFs and the broader CI community grows, it may be possible to foster more tightly integrated activities such as co-development benefitting shared interests and the evaluation of new technologies.

Non-technical challenges influencing CI development

The non-technical challenges discussions focused on the tension between the resources available to support CI and the need to prioritize activities under these constraints. Often, CI improvements appear to receive less attention and associated investment compared to primary science instruments. Additionally, funding agencies frequently prioritize new software development over sustaining and maintaining existing CI, which increases the number of CI migration cycles. These conditions leave LFs having to choose between innovating, upgrading capabilities, and maintaining current solutions while still delivering products and enabling science.

Approaches to enhancing CI workforce and building a community of CI professionals.
Discussions centered around CI workforce were extensive and permeated several workshop sessions. CI practitioners are a mix of domain- and computer-science educated personnel who often need extensive generic and facility-specific CI training. Computer science graduates can be hard to retain in a competitive market with higher salaries in non-science industries. Additionally, their skillsets often do not exactly match LF needs. With no curricula or certifications in the area of CI, finding or even describing the skills needed is challenging. Hiring is also complicated by the fact that enterprise skill sets do not directly translate to the CI environment and some LFs operate in remote locations. To improve hiring and retention, more emphasis is needed on LF career advantages, such as the opportunity to take part in an intrepid science endeavor and better work/life balance than many industry positions.

Diverse backgrounds, siloed CI development, and deployment processes within LFs make it challenging to build a broader CI community. There is no shared description nor clear definition of the skills, interests, and career paths for CI practitioners. Although communication channels exist within each LF, it is hard to catalyze communication and foster community across LFs and CI projects.

Key Findings and Recommendations

As the result of the discussions, workshop participants developed the following findings and recommendations.

Key Findings

- CI challenges: The explosion of large and diverse data sets is driving the decisions LFs make about technologies and software solutions. At the same time, CI is ever-changing and increasing in complexity, and making decisions about and adopting new solutions is complex within an operational LF environment.
- CI challenges: Although scientific networks have improved tremendously, some LFs still struggle to transfer data and access services, especially in remote environments.
- CI challenges: Integration, interoperability, and reuse of cyberinfrastructure solutions could be much improved. There is a natural tension between the community of CI software developers who seek commoditization where applicable and LF missions, which can be highly tailored. Thus, assistance is needed in promoting interoperability via trusted intermediaries.
- CI challenges: There is a natural tension between operations and maintenance needs and those of the end-user, both of which keep increasing over time.
- Collaboration: There are many opportunities to enable collaborations among different LFs or between LFs and CI projects. There is a set of common CI services that could be directly shared or leveraged across facilities. In some cases, lessons learned, best practices, architectures, and technical stacks and their configurations could be shared. In addition, CI should strive to move up the value chain toward more advanced services.
- Collaboration: By fostering collaborative and community efforts, LFs can potentially achieve economies of scale from working together on designing and deploying CI solutions, which in turn may lead to continuity for the solutions used and the ability to deliver a capability beyond the resources of a single LF.
- Collaboration: Fixed budgets for already committed resources and often incompatible timescales are significant barriers to collaboration between LFs and with other CI projects.
- Collaboration: Risk management and mitigation is an integral operational component of an LF that has a direct impact on research outcomes. Additionally, there are still barriers for adopting well-known solutions developed by potential competitors.

- Operational Practices: Management of CI facilities is hard because of mismatches between domain approaches to management and operations in a necessarily interdisciplinary environment (science and computer/engineering).
- Workforce: Hiring, retention, and advancement are particularly challenging at CI facilities because of funding uncertainty, differences between research and industry skillsets and work environments, and a lack of clear promotion paths. However, there are promising activities such as training and professional internships that could be leveraged.
- Workforce: Since hiring from computer science degree programs has not proven optimal, looking at candidates from other backgrounds may be useful.

Key Recommendation Actions:

The workshop participants recommended a number of actions to address the challenges faced by LFs. The recommendations below could be enacted through a combination of community efforts, facility peer interactions, and facility-CI project/platform expertise exchange as well as trusted entities such as dedicated centers.

- Mechanisms for CI discovery and opportunities for sharing of existing solutions, services, and training resources amongst the LFs and CI projects need to be supported.
- A common repository of knowledge about CI best practices, system descriptions, architectures, use cases, and core system tools should be created and made available to the broad community.
- Trusted intermediaries that can help navigate the ever-changing CI landscape, especially when migrating to new solutions need to be funded. Such entities can also assist in science-driven blueprinting of LFs before CI work begins.
- Communication, collaboration, and community-building efforts across LFs and CI projects should be fostered, and the benefits of belonging to a larger society of professionals need to be communicated.
- Research into new and effective methods for incentivizing collaboration and engagement with multiple facilities on joint projects needs to be supported.
- An effort to capture, analyze, and disseminate LF best practices and management techniques and to identify productive engagement opportunities between science, engineering, innovation, leadership, compliance, and other roles needs to be supported.
- Activities geared toward providing affordable training opportunities, helping structure career paths across facilities, recruiting talent starting with the undergraduate level, creating networking opportunities for technical (especially earlier-career) CI staff across facilities, and helping the community to understand the nature of CI as a career need to be supported.

1. Introduction

Overview and Goals

In 2015, the National Science Foundation (NSF) began supporting biennial workshops focused on cyberinfrastructure (CI) for large facilities (LFs). The workshops bring together large facilities and CI projects to share common experiences and challenges, discuss potential collaborations, and identify opportunities for leveraging CI within the two communities.

The 2017 CI for Large Facilities workshop found that "the need for, and benefits of, close interactions, collaborations, and sharing among the facilities and with the CI communities are well recognized, including the sharing of CI related expertise, technical solutions, best practices, and innovations across NSF large facilities as well as research facilities outside NSF (DOE, NIH, NASA, etc.)." Among the recommendations from that workshop was one to "foster the creation of a facilities' CI community and establish mechanisms and resources to enable the community to interact, collaborate, and share" [2017 LF CI report].

The 2019 NSF Workshop on Connecting LF and CI aimed to continue and advance previous discussions, enable the exchange of CI solutions and challenges, and foster CI community building around NSF large facilities. It provided a forum to share ideas and experiences and to prepare for future CI research, development, and deployment that supports cutting-edge science. The major theme of the 2019 workshop was "Connecting Large Facilities, Connecting CI, Connecting People," which emphasized the need to facilitate collaborations among LFs and CI projects and recognized the importance of the CI workforce to LF science missions.

The workshop was publicized at the 2019 Large Facilities Workshop held in April [2019 LF workshop], which included a day dedicated to CI: "Envisioning the Future of Facility Science & Cyberinfrastructure." Participants at the event received a short survey about topics they would be interested in discussing at the September meeting. Based on the survey responses and discussions within the steering committee, a set of initial, high-level workshop goals were identified.

Specific goals of the 2019 workshop included:

- identify CI challenges that facilities face when supporting their science missions;
- explore opportunities and obstacles to collaboration between LFs;
- examine non-technical challenges that influence CI development; and
- develop ideas for enhancing the CI workforce and building a community of CI professionals.

Workshop Organization and Attendees

Steering Committee

The workshop was organized by a steering committee composed of CI experts from LFs and selected CI projects:

- Brian Bockelman, Morgridge Institute and HTCondor project [htcondor]
- Adam Bolton, National Optical Astronomy Observatory [noao]
- Tom Cheatham, University of Utah and Campus Research Computing Consortium [carcc]
- Ewa Deelman (PI and Chair), University of Southern California and CI CoE Pilot [cicoe-pilot]
- Tom Gulbransen, Battelle and NEON [neon]
- Kate Keahey, Argonne National Laboratory and Chameleon [chameleon]
- Marina Kogan, University of Utah

- Dan Stanzione, Texas Advanced Computing Center and the Leadership-Class Computing Facility [lccf]
- Daryl Swensen, Oregon State University and Regional Class Research Vessel [rcrv]

The 94 workshop participants included representatives from the NSF LFs, major NSF-funded CI projects, selected cloud providers, and NSF staff. Every effort was made to have at least one representative from each LF [lf_list, Appendix C]. For facilities with a distributed collaboration, representatives for the various LF platforms were invited. This was the case for the National Hazards Engineering Research Infrastructure [nheri], which includes a number of experimental facilities as well as different class vessels that make up the Academic Research Fleet [arf]. The list of attendees is found in Appendix A.

Workshop Structure and Activities

The major topics of the workshop built on discussions, findings, and recommendations from previous workshops and meetings. The agenda for the workshop was defined by the steering committee with community input. A spring 2019 survey advertised to LF workshop participants that attended CI day helped define the overall goals of the workshop by soliciting ideas for discussion topics and suggestions about people to invite to the workshop. A second survey sent to workshop participants prior to the meeting helped define particular panels and breakout sessions. The participants were also offered the opportunity to give a short "lightning" talk about their work, which resulted in 12 presentations. Finally, the workshop included a guided activity in which participants were divided into small groups and asked to answer two questions: 1) What are the most significant challenges faced by LFs or projects? 2) What are the most important problems a Cyberinfrastructure Center of Excellence could solve? The groups discussed the questions and were given the opportunity to share their answers verbally with all groups and to contribute their answers to the workshop notes.

Kate Keahey (ANL) and Rafael Ferreira da Silva (USC) lead the Guided Activity Session.

Meeting participants were given the opportunity to provide input on workshop notes taken using Google Docs. Selected participants (Appendix A) were also asked to take notes during all the sessions to ensure that the discussions were captured. The raw, read-only notes are linked to the workshop agenda page (https://facilitiesci.github.io/2019/agenda.html).

To obtain a quick pulse of the community, workshop participants were asked to send a survey to their teams to gain information about knowledge sharing, incentives to following a CI career path, and feelings of belonging to the broader professional CI community. The analysis of the results is included in Section 7.

The workshop also started a community-building effort by requesting participants to provide "CI calling cards" that included their professional information, pictures, and responses to questions about their recent CI frustrations and successes, both technical and non-technical. These calling cards greatly facilitated group and one-on-one discussions and are included in Appendix D.

This report is organized around the main topics of the workshop rather than the program components. Findings are formatted in *italics,* and recommendations are underlined. All workshop materials, including presentations, notes, photographs, and calling cards, are available on the workshop website: http://facilitiesci.org.

2. Theme I: Identifying cyberinfrastructure challenges that facilities face

Prior CI for LFs workshops identified the technical challenges of keeping up with acquisition; processing, and delivery of vast, heterogeneous, and dynamic data; tracking and managing technological changes; employing automation to streamline data processing and operations; developing well-informed and robust CI; and sustaining CI solutions over time. Although the various challenges were discussed throughout the meeting, the breakout session "What are the CI challenges that need to be addressed in the next 5 years to support LFs?" was dedicated to these discussions. In this theme, we highlight some of the challenges discussed in greater detail at the workshop.

The explosion of large and diverse data sets is driving the decisions LFs make about technologies and software solutions. The growing complexity of sensors and instrumentation, longer lifespans of LFs, and technological advances that make it possible to cheaply acquire and store data is leading to an overwhelming explosion in the volumes, acquisition rates, and heterogeneity of data. The data explosion not only affects short-term storage and long-term archiving but also decisions LFs make about technologies and software solutions. Many LFs must balance their budgets with priorities of long-term archival and real-time delivery of instrument and pre-processed data. For example, LFs that rely on relatively slow and expensive satellite uplinks to move data from instruments to data processing or archiving centers, such as IceCube and the Academic Research Fleet, must balance uploading high priority science data with quality of life issues such as internet access for the staff at a facility or on a ship.

Although scientific networks have improved tremendously, some LFs still struggle to transfer data and access services. Many LFs feel that there are times when networking can still be challenging in modern, heavily connected environments. Problems can occur when performing high-bandwidth transfers between data centers across the continent or between online storage and backup facilities. An interesting discussion point was that bandwidth is not the only concern for the more sporadically connected facilities. There is also a need to "always be connected" to use online collaborative tools such as Dropbox, Google Docs, Microsoft One Drive, or when dealing with software licenses such as the Adobe software suite that requires an online network connection.

Daryl Swensen, Oregon State University, reporting on the breakout session "What are the CI challenges that need to be addressed in the next 5 years to support LFs science missions?"

CI is ever-changing and increasing in complexity, and making decisions about and adopting new solutions is complex within an operational LF environment. Software development is also an area of concern for many LFs. Because many LFs have longer lifespans, the challenges most often brought up by participants related to long-term stability, supporting legacy code bases, and allowing scientists to use a never-ending, constantly changing set of modern software tools. For LFs to operate smoothly, they must perform thorough requirement analysis while managing expectations from a funding agency and the community they serve. However, this requirement analysis often needs to be revisited many times during the lifetime of the LF. Several attendees also mentioned that compound interest on technical debt is expensive, and keeping and supporting older codebases is usually not a good option. Another software-related consideration is whether to use academic-grade software or commercial solutions, both of which could eventually bring prohibitive cost. The former solution might go away or become unsupported while the latter might lead to vendor lock-in and has its own sustainability challenges. Additionally, many LFs stated a desire to increase their use of machine learning algorithms for recognizing complex data patterns. However, state-of-the-art machine learning is available primarily in modern, community-driven software tool chains that present transition barriers to LFs.

Integration, interoperability, and reuse of cyberinfrastructure solutions could be much improved. One challenge discussed at the workshop was that facilities typically address CI challenges independently of each other and develop custom solutions in an uncoordinated manner. The result is that they are re-inventing similar solutions and miss opportunities to leverage work and knowledge. LF lifespan and different project timelines can make it difficult to identify and commit to long-term CI partnerships.

There is tension between operations and maintenance as well as the needs of the end user, the latter increasing over time. LF mission focus and the need to mitigate long-term risk often result in the desire to invest in in-house software—resulting in a plethora of bespoke systems—rather than adopting existing solutions or working collaboratively across LFs and large CI projects. Timelines and budgets under which LFs operate also may drive them to pursue expedient solutions rather than step back and examine potential new technologies and solutions. The same pressures may limit LF willingness to seek and maintain collaborations that can potentially increase CI reuse and new CI adoption.

Patrick Brady, University of Wisconsin Milwaukee, gives a lightning talk about "SCiMMA: Scalable Cyberinfrastructure to support Multimessenger Astrophysics."

Recommendations

The workshop found that for LFs to develop, adopt, and maintain robust CI requires planning that is flexible enough to allow agility in moving forward in incremental steps over the LF lifespan. Planning, including quantifying the potential benefits and the cost of making changes, is important because of the difficultly in making changes to Major Research Equipment and Facilities Construction projects (MREFCs) in progress even when the technology landscape is continuously and rapidly changing. During the conceptualization phase, it is especially important to bring in CI expertise, people with a frame of reference for operations, and people with a broad overview of current robust CI capabilities. During the LF operational phase, it is critical to re-integrate with the broader CI community to leverage each other's strengths and develop approaches to change management. There is a need to create opportunities for CI discovery and sharing of existing solutions, services, training resources amongst the LFs as well as CI projects.

A somewhat obvious, yet important, recommendation is to facilitate the use of best practices and effective processes. Creating a common repository of knowledge about CI best practices, system descriptions, architectures, use cases, and core system tools—e.g., a taxonomy of logical architectures for LFs—is desirable. The CI community should work towards a set of guidelines on creating well-developed documents that capture architectural details, rationale, and alternatives studied before reaching the decision on CI. The knowledge repository could also include information about common software and services, data formats and ontologies, data curation and preservation techniques, and reproducible results. The ability to access community wisdom will become increasingly important as the principles of findable, accessible, interoperable, and reusable (FAIR) [fair] data are adopted by more scientific communities. The attendees were somewhat divided as to whether the knowledge base should also include evaluations and recommendations of some tools over others. Finally, a yearly, focused gathering of CI/LF senior architects may be a very effective way to complement (but not replace) the common repository as discussed above.

There was a lot of interest and recommendations for a trusted entity, collection of trusted entities, or Centers of Excellence. In general, the agreement was that the ever-changing and increasing complexity of CI raises the need for trusted intermediaries to help navigate the CI landscape. These intermediaries would assist in science-driven blueprinting of LFs before CI work begins, provide a mechanism to "matchmake" between LFs and CI projects, and decide when the partnership works for both or could be applied across projects. A more complex recommendation was for the trusted entity to help clarify the distinctions, possibly case-by-case, between common/commodity CI solutions and those that are domain-specific and to also clarify the distinction between wants and needs for facilities.

Finally, commercial cloud solutions provide a number of capabilities (computing, archival storage, etc.) that can contribute to the LF science mission. However, the published cloud costs often appear cost prohibitive. The community could benefit from an effort to work with all cloud vendors to explore whether the commercial cloud can contribute to LF and CI in a cost-effective manner and provide the support that the science community needs. A potential area of exploration, which would potentially increase sustainability of facilities and researchers who create scientifically valuable data sets and algorithms, are new business models that incentivize cloud vendors to share their compute revenue with the creators of the content/algorithms subscribers use.

Margaret Johnson, University of Illinois at Urbana Champaign, gives a lightning talk about "Enabling Multi-Instrument Pixel-Level Science with A High Throughput Computing, Data"

3. Theme II: Exploring the opportunities and obstacles to collaboration between LFs

Collaboration is essential to support and strengthen scientific progress. Bringing additional expertise and resources together can enable research endeavors that were not otherwise possible. On the other hand, when not properly managed, collaborative efforts may raise potential risks, challenges, and inefficiencies. Several discussions about opportunities and challenges to collaboration between LFs took place during the two-day workshop, including the panel discussion "Shared CI Services: Opportunities and Challenges." In this section, we summarize and highlight findings and recommendations for these opportunities and barriers to collaboration.

There are many opportunities to enable collaborations among different LFs or between LFs and CI projects. Although different facilities target different scientific problems from distinct science domains, there are a set of common CI services that could be directly shared or leveraged across facilities or about which lessons learned, best practices, architectures, and technical stacks and their configurations could be shared. Notably in the Participant and Practitioner Survey (see Appendix for full results), half of the respondents selected somewhat likely or very likely when asked about the likelihood of their facility/project being willing to contribute expertise to a shared services efforts. Potential common CI services include compute, storage, authentication, data discovery, data archive, disaster recovery preparedness, deployments, maintenance, and networks.

Current models for sharing resources among facilities include: (1) organizational (e.g., ITIL/COBIT model [itil] for shared IT service & governance), (2) mission-driven (e.g., DOE-funded CI centers), (3) vendor-oriented (e.g., commercial clouds), (4) partnership-based (e.g., partnership agreements with major NSF-funded CI centers), (5) collaborative (e.g., facilities may share CI with one another), and (6) ad hoc (e.g., people who occupy similar positions at different LFs informally communicate or meet at common conferences such as the forum AskCI [askc] to share recent experiences and opinions). The first three models are currently achieved via internal organization and economic incentives but have very limited scope. Partnership-based agreements with CI centers have demonstrated an impact on domains such as cybersecurity (Trusted CI [trustedci]) and science gateways (SGCI [sgci]).

Panel on Shared CI Services Opportunities and Challenges, Moderator: Adam Bolton (National Optical Astronomy Observatory). Panelists: (from the left) Pamela Hill (National Center for Atmospheric Research), Von Welch (Indiana University and Trusted CI), JJ Kavelaars (National Research Council of Canada), and Michael Zentner (University of California San Diego and Science Gateways Community Institute).

By fostering collaborative and community efforts, LFs can potentially achieve economies of scale from working together on designing and deploying CI solutions, which in turn may lead to continuity for the solutions used and the ability to deliver a capability beyond the resources of a single LF. Collaboration can be achieved by sharing expertise, connections, and efficient processes that allow LFs to focus on their core competencies. For instance, NOAO [noao], LSST [lsst], and GEMINI [gemini] projects share several CI services, mostly focused on computing and data management, across their institutions. In addition to technology components, sharing competence, community, knowledge, and awareness also aggregates high value to the collaboration and may lead to improved project sustainability, funding acquisition, and workforce development. Furthermore, commonality in architectures in what are often niche academic spheres, can drive a tool or service in a direction that benefits all and develops a wider, more skilled community. Although this can occur for both commercial or open source tools and services, it is particularly likely to happen with open source resources for which LF CI staff are the primary developers. To establish such collaborative efforts among the diversity of scenarios for LFs and CI projects, several challenges need to be addressed, as discussed below.

Risk management and mitigation is an integral operational component of a LF that has a direct impact on research outcomes. Although outsourcing less specific and critical CI services (e.g., by leveraging other CI systems and platforms for computing and data management) may improve service robustness and availability, there is a perceived risk when relying on external software, hardware, and/or connectivity

services. For instance, moving large data sets over the wide area network may be challenging due to low network connectivity and may also raise security issues. Accessing shared computing platforms requires predictable job execution and turnaround times as well as the ability to support complex model configurations (e.g., multiple job steps, in-situ analysis, data assimilation, machine learning components, different software environments, etc.).

There are still barriers for adopting well-known solutions developed by potential competitors. For instance, a common approach in CI software development is to redevelop solutions rather than tailoring community software to specific needs. This approach is typically motivated by the need to host solutions developed in-house (software development path is customized to the LF need), lack of documentation, and/or long-term sustainability. The lack of formal common forums for collaboration and sharing of experience also inhibits competence sharing.

The timescales at which LFs operate is another hindrance to collaboration. It becomes incredibly complex to share information or collaborate with quickly changing technologies, especially when processes and budgetary structures do not support refactoring, evolution, sharing, interoperability, planning upgrades, and new deployments. Those factors are further complicated by the fact that simultaneous old and new deployment periods are inevitable during the upgrade of an operational facility.

Finally, there are both monetary and physical barriers that impose constraints on collaborations. CI practitioners working in remote environments such as ships or sites that are beyond connectivity simply cannot regularly communicate with others. Additionally, LFs have limited resources that are already allocated to serving their target users. Developing new relationships, forums, and other vehicles for collaboration is costly in the immediate timeframe, even if it will reap a multiplication in long-term rewards. Despite the above discussed willingness to share expertise, respondents to the Participant and Practitioner Survey were ambivalent when asked about the likelihood of their facility/project being willing to contribute resources (provide storage, compute services, personnel). And when asked what obstacles they foresaw in implementing a shared services effort, the largest portion of respondents (40.63%) indicated funding and personnel as the main obstacles.

Recommendations

Given the above, there is a clear opportunity and long-term value for the community and funding agencies to put in place mechanisms that enable collaboration. Excellence in collaboration requires trustworthy people, processes, and technology. Increased risks in collaborative projects can be mitigated by utilizing existing trusted relationships and working to actively develop new ones.

In the immediate term, there are some relatively simple activities that could be undertaken. For instance, sharing existing technical architectures in a form that makes them discoverable and reusable would allow LFs to recognize which other LFs might have relevant expertise to share. Additionally, it could inspire consideration of similar architectures while benefiting from lessons already learned. Similarly, the sharing of expertise regarding open source licenses, community data standards, and software development best practices were all voiced by the workshop participants as simple but tangible potential products of value if shared in a useful manner.

The results of the Participant and Practitioner Survey showed a willingness (within given caveats) to share expertise alongside the acknowledgement and calling out of the obstacles of limited funding and personnel time. Thus, it is plausible that tangible funding should be made available specifically to support collaborative sharing efforts. However, more research is needed to carry out a cost benefit analysis of sharing models before such a recommendation be made.

Lastly, as a longer-term effort, external support structures could foster and facilitate increased collaboration. For instance, external support could facilitate and foster possible community collaborations that appear to have the long-term potential to justify initialization cost. These collaborations will likely fall along a specific CI theme. This support could include maintenance of mailing lists, organization and hosting of monthly telecons, planning, recording, and distributing webinars that share LF architectures and lessons learned.

Other instantiation of external support could include a facilitator curating a special issue publication showcasing technical architectures (perhaps themed according to a stage of LF data lifecycle) or a publication surveying emerging technologies. Whether the latter would be written solely by external evaluators with the technical expertise to do so, or by LF staff could be domain dependent. Ultimately, these publications could serve as a source of collated and reviewed material that saves each LF from carrying out its own in-depth evaluation of every new technology that appears. Clearly, these evaluations would need to clearly detail the use case context of the evaluation.

Finally, the workshop attendees recommended that a team dedicated to short-term projects involving multiple LFs (and requiring LFs to provide their own effort) would be highly effective. This final recommendation, however, is likely the most expensive.

4. Theme III: Examining non-technical challenges that influence CI development

Several non-technical issues and challenges that influence CI development for LFs and, more broadly, for large NSF CI projects, were identified during the discussions at the workshop. This section captures those discussions and recommendations from participants about how to overcome those challenges.

There was also a significant amount of discussion around models for a trusted entity or set of entities that could facilitate or nucleate community efforts, facility peer interactions, facility-CI project/platform expertise exchange, or provide services to the CI community. The example of the collaboration between the Cyberinfrastructure Center of Excellence Pilot [cicoe-pilot] and NEON [neon] was discussed as one possible type of a one-to-one engagement. The participants also shared ideas about other useful tasks that the community, LFs or other entities could perform to improve the LF CI development. Among them were:

- perform an assessment of all/many LFs to identify common needs and problems,
- evaluate new technologies and help LFs understand which are applicable to their use cases,
- provide training and recommendations while facilities stand up the hardware and software,
- maintain expertise in specialized areas (e.g., Internet of Things, workflows, data modeling, data archiving) where individual facilities cannot afford to do because of episodic rather than constant needs,
- provide opportunities for engagement between LFs and CI providers at key points in the operational timeline to meet needs for general consulting or for specific solutions/services,
- develop expertise and consulting around standards (not act as an auditing body though) e.g., evaluate NSF mandated proposal requirements with respect to sharing CI for LFs, and
- provide an ‘army’ of skills.

Breakout session on "What are the non-technical issues that influence CI development and how they can be collaboratively addressed?" The session was led by Susan Sons (Indiana University) and Douglas Thain (University of Notre Dame).

The management of CI facilities is difficult because of mismatches between domain approaches to management and operations in what is necessarily an interdisciplinary situation (science and computer/engineering/CI). Some of the challenges in operational practices are enumerated below.

There is a tension between adopting new technology/innovation/improvement and maintenance/feasibility/sustainability of CI. One of the operational challenges for large CI entities is to develop a coherent strategy for achieving balance between sustaining what has been developed vs. engagement in new activities with the consideration that resources might also be needed for urgent activities. There is often a lack of long-term planning for ensuring these different needs are kept balanced and for dealing with legacy infrastructure.

There is a lack of clear understanding about the ideal organizational structure within large CI facilities. The teams are often highly distributed and frequently exhibit "split agency" characteristics, meaning there is a mismatch between individual and organizational objectives resulting in conflicting demands and expectations. Sometimes, CI teams serve multiple PIs with conflicting needs and desires. Another concern is balancing the organizational structure to cater to domain science expertise and computing/CI expertise.

Staffing skills and knowledge transfer are a challenge faced by most facilities. Often, large CI facilities are heavily matrixed and are trying to maintain a small staff with many specialized skill sets. When turnover occurs, knowledge is lost. This is especially problematic when a developer leaves behind systems that were created using specialized knowledge and unique skills. How to manage these systems in a way that reduces loss of maintainability is a significant challenge. The lack of communication among technical staff and between technical and non-technical staff also impedes smooth knowledge transfer during successions. Insufficient tacit knowledge capture and lack of automation in knowledge capture are serious challenges for long-lived CI projects and LFs.

Recommendations

There needs to be an effort to analyze and distill best practices and management techniques, and for identifying productive engagement opportunities between science, engineering, innovation, leadership, compliance, and other roles that would be instrumental for LFs.

There needs to be support for research focused on successful staffing models (science vs. computer/engineering/CI) and resource management practices, and a way to share those outcomes with NSF, LFs, and large CI projects and organizations. It would be beneficial to create a body of best practices around well-integrated CI engineering organizations within science organizations.

There is a need for creating trusted entities that can follow relevant technology trends, evaluate new solutions in the context of LF CI, make recommendations, and provide training.

5. Theme IV: Developing ideas for enhancing the CI workforce and building a community of CI professionals

Discussions about enhancing the CI workforce and building a community around CI permeated most workshop sessions. Developing and growing a community of CI practitioners available to support the science missions of LFs was identified as one of the major challenges faced by workshop participants. The word *community* can be difficult to define, particularly with no unifying entity to establish an identity for the community or to govern its actions. Workshop participants agreed that a community is a group of people who share a common language and common function or purpose, but the boundaries are not always clear. For example, a university, LF, or organization like XSEDE [xsede] could be a community, but so can a department within a university, a professional society spanning multiple organizations, or a group of people with similar interests from a variety of different locales.

Communities should be venues to share expertise and experience, despite the challenges in drawing strict lines around communities. They should have a centralized way to communicate and exchange ideas. People within communities should see mutual benefit in belonging and should be self-selecting. A Cyberinfrastructure Center of Excellence could help people self-organize into the much-needed community of CI practitioners.

Today's CI practitioners may have found their positions at LFs via an internship program (e.g., NCSA Cyberinfrastructure Professional Intern Program [cip, mate]) or a contract position where they were assessed for fit before being offered a permanent position. Of those who receive a permanent posting, workshop participants estimate about two out of three stay.

Workshop participants found many challenges associated with recruiting:

- Compared to industry, some LFs have requirements or working conditions that may make them less appealing. For example, working at the South Pole requires tolerance to cold and isolation, and working on research vessels may exclude candidates who suffer from chronic medical conditions if they need continual specialized care. Some LFs require employees to work remotely for long periods of time, away from home and family.
- Typical computer science curriculums may not adequately prepare students for work in an LF.
- Graduating students may not even be aware of LFs. Campus recruiting fairs may not think to include LF representation, and career services' counselors may not be aware of LF opportunities.
- Job descriptions may not accurately represent the role because the institution's human resources may require that postings fit titles and templates that suit their needs, rather than the needs of the

hiring manager. For example, some participants said they had to name positions with more run-of-the-mill titles, such as software developer, when the role is substantially more nuanced than that. This can lead to misunderstanding about the requirements of the job and frustrate the hiring process.
- Because LFs tend to operate differently from their parent institution/university, hiring managers may have difficulty in explaining or justifying hiring, compensation, or promotion needs to a human resources department that primarily deals with standard university hiring and retention.
- Recruiting international employees may prove difficult or impossible due to the institution's export control practices or reluctance to, or difficulty with, managing visas.
- Salaries are not competitive with those offered in industry. Google, Microsoft, and other high-profile tech companies can pay far more than an NSF budget allows.
- Funding uncertainties such as limited term cooperative agreements make it hard to hire and retain quality CI professionals.

Retaining staff is also a challenge due to a number of factors:
- Career paths are not well defined. There is no common set of job titles and descriptions, paths for attaining raises and promotions, or sets of necessary certifications or qualifications. Instead, people tend to "fall into" CI roles, which makes it harder to justify positions or the value of the individual employee (and subsequently his/her raise or promotion) at performance review time. Because the CI practitioner role(s) has not been formalized across the market, institutions/universities are hampered in providing advancement, development, and proper compensation.
- CI practitioner roles tend to be filled by individuals of the same demographic. Few women and people of color appear in these roles. Those who do may be less likely to stay if they feel isolated from or socially different from their co-workers or perceive imbalances based on race and sex.
- Many teams at LFs are geographically distributed. They may work at different time zones, speak different languages, and communicate less frequently or via less rich means (i.e., not face-to-face). Although this diversity has many benefits, it also proves challenging for coordination, communication, and establishing a sense of shared identity and community among team members.
- Some LFs require employees to travel for extended periods of time, and science goals may need to be addressed on accelerated schedules. Both can make it difficult to carve out time for employee development and training.
- Sometimes there is a disconnect between the scientists and CI practitioners. They may not share the same language and will necessarily approach problems from different perspectives. This may make common understanding, and consequently achievement of goals, challenging. To avoid frustration, time should be taken to either cross-train, find individuals with a blend of these skills, or foster communication and understanding.

Although there are many challenges around this workforce issue, there are many things that work well and can be further exploited to improve the situation. For example, hiring managers can appeal to work/life balance, which workshop participants felt was more favorable than in the private sector, or to the science mission which, to some candidates, would make working at an LF meaningful and enriching. Many training opportunities already exist, including Open Science Grid's user school [osg], TrustedCI training, and the Linux Cluster Institute. Some networking opportunities exist as well, including the workshop that this report discusses, the Campus Research Computing Consortium (CaRCC) [carcc], and the Coalition for Academic Scientific Computing (CASC) [casc]. Additionally, LFs such as the Academic Research Fleet [arf] have a successful MATE program [mate].

Additionally, one could create a trusted entity or set of entities to help:
- Research and provide a repository of information and resources around developing community and the CI practitioner workforce, including:
    - mailing lists around topics of interest,

    - lists of sub-communities and their annual meetings,
    - best practices, guides, and templates on matters of community building, hiring practices, employee development, etc., and
    - hosting a CI job board.
- Serve as a connector to bring people together by listening to and knowing the larger LF community so that members can make meaningful connections with and receive help and guidance from peers. This could also involve serving as a "birds of a feather" space where people can congregate virtually or in person and self-organize into subcommunities.
- Create opportunities for engagement and knowledge sharing, such as hosting a version of StackExchange for CI practitioners; facilitating or helping establish mentoring programs; offering or collaborating on the development of training on all aspects of CI; facilitating or helping establish an exchange program where CI practitioners work at other LFs for a time; offering other networking and knowledge-sharing events.
- Work with universities to make students and career services more aware of job opportunities in LFs.

Recommendations

There needs to be an effort to help facilities meet workforce requirements by providing affordable training, creating career paths across facilities, recruiting talent starting at the undergraduate level, and helping the community understand the nature of CI as a career. One recommendation is to focus on hiring domain science graduates to work in computing/CI areas. It is also important to recognize that research CI professionals often remain in their jobs even if the pay is not competitive because they enjoy the work and science. It would be beneficial to create an “academy” for improving computing/CI skills that are broadly applicable for LFs.

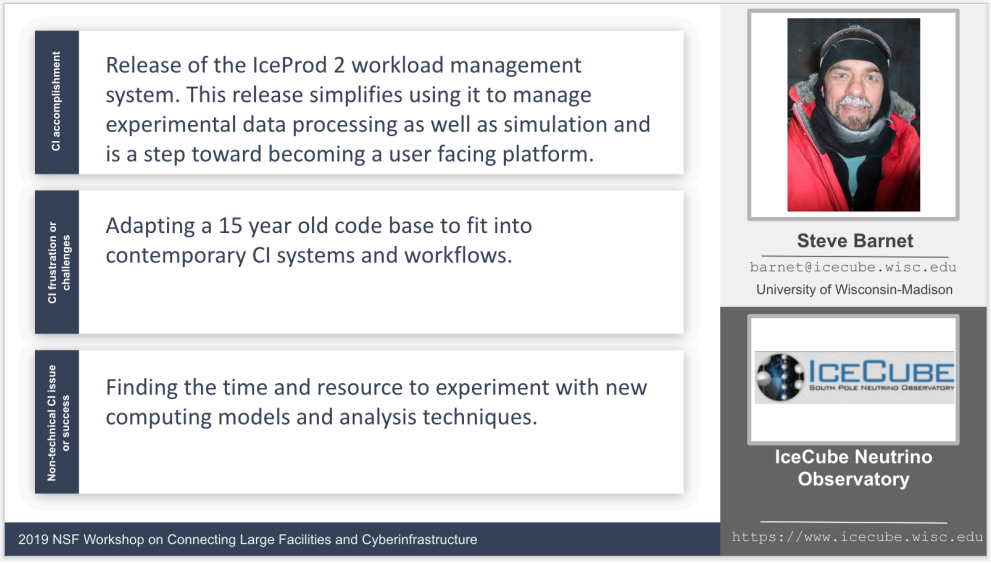
A trusted entity could also provide staff training on emerging solutions that a specific LF does not have time to provide. Articulating career paths for CI professionals is also important. This could be done by interfacing with undergraduates to make career paths known and by talking to career services at universities. It could also be beneficial to research factors that cause people to stay or leave CI jobs, potentially designing and analyzing exit surveys. Exploring avenues like cross-facility job postings or offering CI experts the opportunity to work at other facilities for a period rather than leaving for industry, could also help. It may be beneficial to develop an internship funnel for under-represented communities and facilitate Research Experiences for Undergraduates (REU) programs related to CI.

*Panel on Workforce Development and Retention, From the left: Sharon Broude Geva (University of Michigan); Tom Cheatham (University of Utah, panel moderator), Frank Wuerthwein (University of California San Diego and CMS), Rachel Adams (UC Boulder), and Jim Rosser (*IODP JRSO).

6. Cyberinfrastructure Calling Card Summary

Prior to the workshop, participants were asked to complete a CI calling card that was used to help participants meet each other virtually before the workshop and to facilitate conversations about topics of mutual interest between participants during the workshop. The calling cards asked participants to briefly comment on three topics:

- a CI accomplishment they achieved over the past year,
- a CI frustration or challenge they faced over the past year, and
- a non-technical CI issue or success they would like to share with workshop participants.



An example of a CI Calling Card.

Participants submitted 61calling cards, which were analyzed after the workshop for common themes or patterns. Calling card comments were grouped into two main categories: (a) accomplishments and

successes and (b) frustrations, challenges, or issues. These were further subdivided into themes that manifested during the analysis process.

Themes within comments about accomplishments and successes:

- *Deployment of something new (e.g., systems, services, infrastructure)*. Examples included deploying clusters around the globe or on research vessels and creating system assessment tools and metrics that allow reporting to stakeholders.
- *Improvement, expansion of an existing system, service, or infrastructure*. Examples included expanding storage, automating cyber-physical experiments, and provisioning 10 Gbps fiber to telescopes in remote locations.
- *Continuing to do good work throughout the year*. This was notable because often day-to-day activities may not be characterized as achievements due a definition of accomplishments as novel and set apart from the day-to-day. However, duties done well are vital to the successful operation of any type of organization, including large facilities. Examples included ensuring data quality is assessed and monitored; finishing construction and launching on time and within budget; and balancing the needs of stakeholders while addressing challenges.
- *Communication-oriented activities*, specifically around bringing people together, were mentioned frequently. Examples included helping distributed units of a large facility connect and feel more like a network, partnering with peer organizations to make improvements, participating in related communities (e.g., EarthCube [earthcube], ESIP [esip]) to advance shared goals, building effective teams, and facilitating collaboration between researchers and developers.

There were other types of successes mentioned that did not fall into any of the above themes, such as instituting an efficient procurement process, meeting the needs of funding partners, successfully meeting a standards base (e.g. Open Data Act, ITAR) or applying one, and leading development efforts.

Themes around frustrations, challenges, or issues:

- *Frustrations around data* included those around sharing, disseminating, and accessing data; data management and curation practices; and the impact that increasing data acquisition rates have on things like storage and management.
- *Budget woes*, which include scaling up without the budget spinning out of control as well as mismatches between operations timelines and funding cycles that may negatively impact the ability to respond to evolving technologies.
- *Technical issues*, which can be further divided into:
    - *Planning, systems analysis, and design*. For example, learning there is more to manage (e.g., support, migrate, refactor) or more is needed to satisfy needs than was originally anticipated or planned for.
    - *Old vs. new technology*, such as responding to evolving needs or keeping older systems up-to-date with current technology.
    - *Heterogeneous technology*, such as integrating different technologies based on differing needs; the need for specialized or customized tools because one size does not fit all; or working with researchers that use different platforms, operating systems, software, etc.
    - *Shared services and distributed services*, such as the need to gain the attention of shared resource providers; or managing the details of distributed storage allocations.
- *Communication-oriented challenges,* including changing culture and mindsets around a variety of things (e.g., motivating scientists to publish data, encouraging operational thinking, redefining value); contending with conflicting or changing needs of stakeholders; and communicating with administration on mission focus or technology costs.
- *Workforce-related challenges* include a sense of disconnect between CI professionals and scientists or needing to blend the two skill sets; recruitment and retention of professionals in a competitive technical market; or overtaxed staff due to multiple funding obligations and reporting lines.

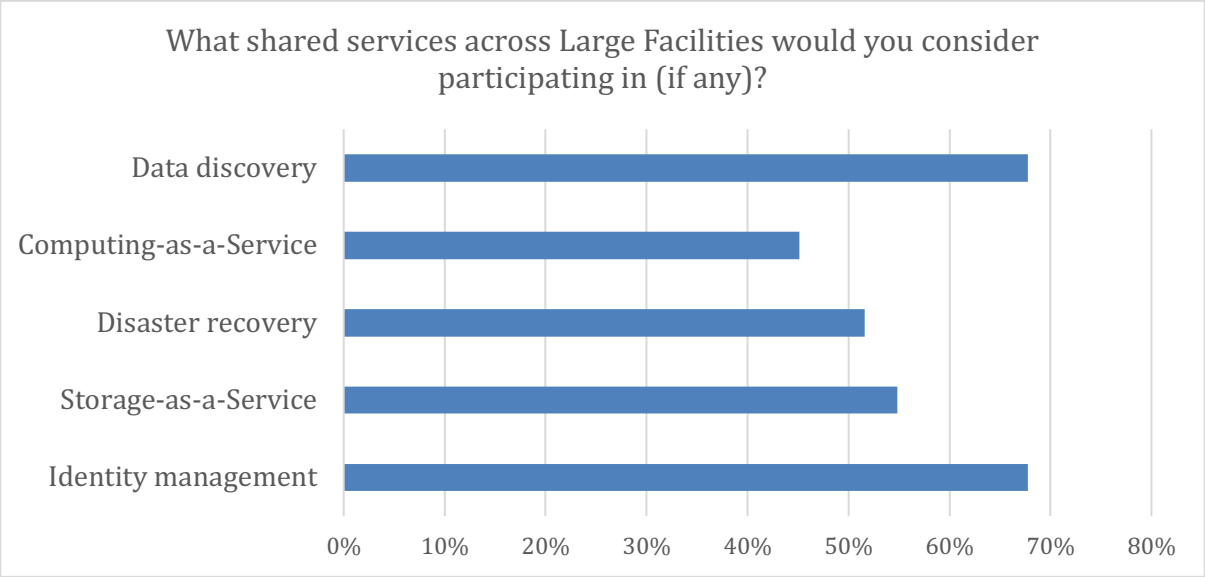
Other challenges, frustrations, or issues mentioned that did not fall into one of the above themes included desires for more information on best practices; limited bandwidth and reliability of internet connectivity on research vessels; and complying with multiple disparate standards.

7. Participant and Practitioner Surveys Summary

Participant Survey

Thirty-two workshop participants responded (43% response rate, since we did not send the survey to NSF staff) to a pre-workshop survey that focused on opportunities for collaboration and sharing expertise across LFs.

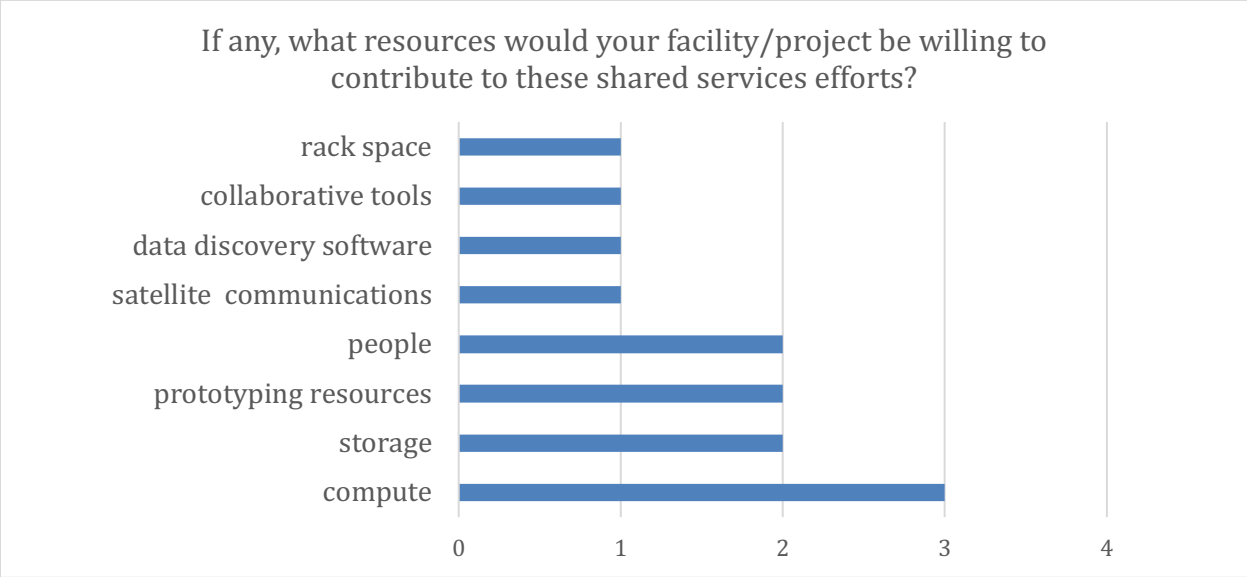
Respondents indicated LF-shared services in which they would consider participating. The survey provided the following options (multiple choice): Identity management, Storage-as-a-Service, Disaster recovery, Computing-as-a-Service, Data discovery, and the ability to write in additional services. The graph below shows that many participants saw the benefits of sharing these services.



Respondents also wrote-in additional services they would like to potentially participate in. These "write-ins" included: shared personnel services to support contracting, training and workforce development, diversity, organizational and policy aspects, satellite communications, global integration of computing and data, global content delivery system, and resource scheduling.

Respondents were asked about the likelihood of their facility/project being willing to contribute resources (provide storage, compute services, personnel) to the shared efforts. For the response, they were offered a 5-point Likert scale (from 1: very unlikely to 5: very likely). For the willingness to contribute resources, the response mean was 2.94 and median 3, indicating that respondents are rather ambivalent about committing resources to a shared service.

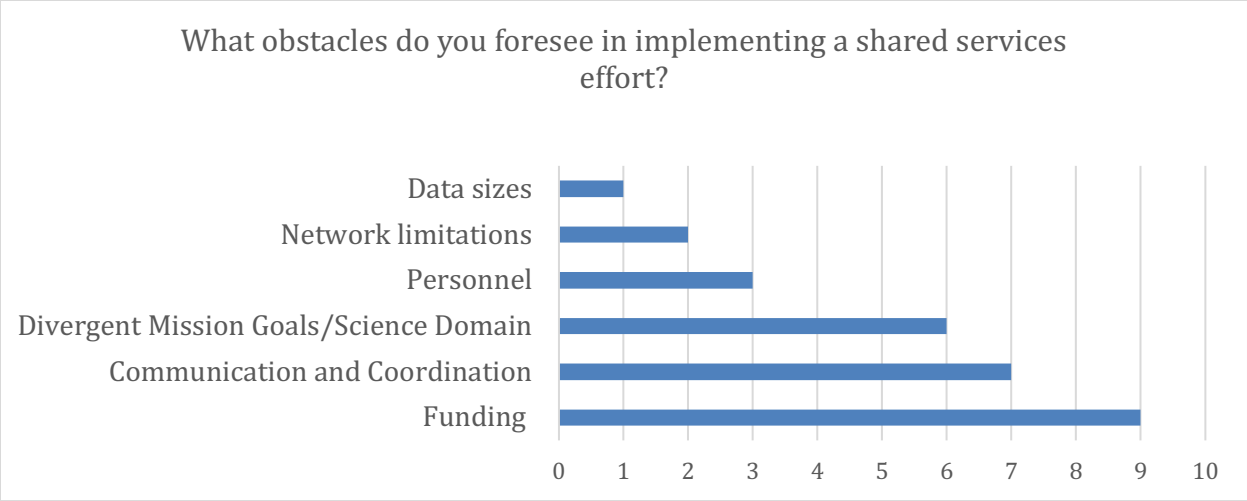
When asked which resources their facility/project would be willing to contribute to these shared services, most respondents did not answer the question or indicated the inability to share any/many resources. Only eleven respondents indicated a few salient types of resources their organizations would be willing to share:



On the other hand, when asked about the likelihood of their facility/project being willing to contribute expertise to the efforts of shared services, the respondents were much more optimistic, with mean of 3.69 and median of 4 on the 5-point Likert scale. Overall, respondents were much more willing to contribute their expertise to a shared service rather than resources. The high median indicates that half of the respondents selected somewhat likely or very likely, signaling a high level of buy-in. This is well-illustrated by this participant quote: "I would think an expert-exchange program would be excellent – if nothing else a forum for knowledge exchange."

When asked what specific expertise their facility/project would be willing to share, the respondents indicated many diverse types of expertise, including identity management, network engineering, cybersecurity, storage, statistics, data science, high-performance computing, remote, shared CI, standards, restful service layers, user management, cloud computing, replicability of experiments, user access, and personnel.

When asked what obstacles they foresaw in implementing a shared services effort, the largest portion of respondents indicated funding as the main obstacle, at 29%. The other salient categories are shown in the graph below.



One respondent stressed the lack of willingness to collaborate as the main obstacle and pointed out its relationship to the life cycles of LFs: "None other than the willingness to collaborate. In general, large facilities projects are very insular in their initial phase because they are totally inward focused on getting the job done and perceive everything shared as a risk. So, it is a matter of control, and risk mitigation. As they mature, they tend to be more likely to collaborate. Except for the fear of 'their money' being spent on a collaborative effort that they do not control. So, a fear of losing funding and control are the real obstacles."

A large proportion of participants indicated online platforms like discussion forums and chatrooms (Slack, Microsoft Teams, etc.) would be helpful for discussing CI-related issues (28.13%). One respondent illustrated the hopes for such an online space quite well: "single point (website?) for discoverability of current CI providers, initiatives, and solutions." Other popular options included continuing annual CI meetings (6.25%) and a combination of in-person meetings and digital platforms (9.38%). Another participant indicated that the platform is less important than the motivation to contribute: "Any platform will work, but 'helpful' will happen only if people use the platform. Infrastructure and shared service discussions require domain expertise and IT expertise and topical abstraction up and down, requiring regular and repeated discussion at many levels and probably multiple platforms. How do we motivate people to engage in the conversation?"

Next, respondents were asked about shared workforce development activities across NSF LFs. They were first asked what aspects such shared efforts would need to have for participants to consider participating in them. Funding or attractive costs were again the most prominent category (12.5%). Other salient aspects of the shared workforce development effort included activities that enhance the ability for LF employees to share knowledge, and HR topics/retention for supervisors.

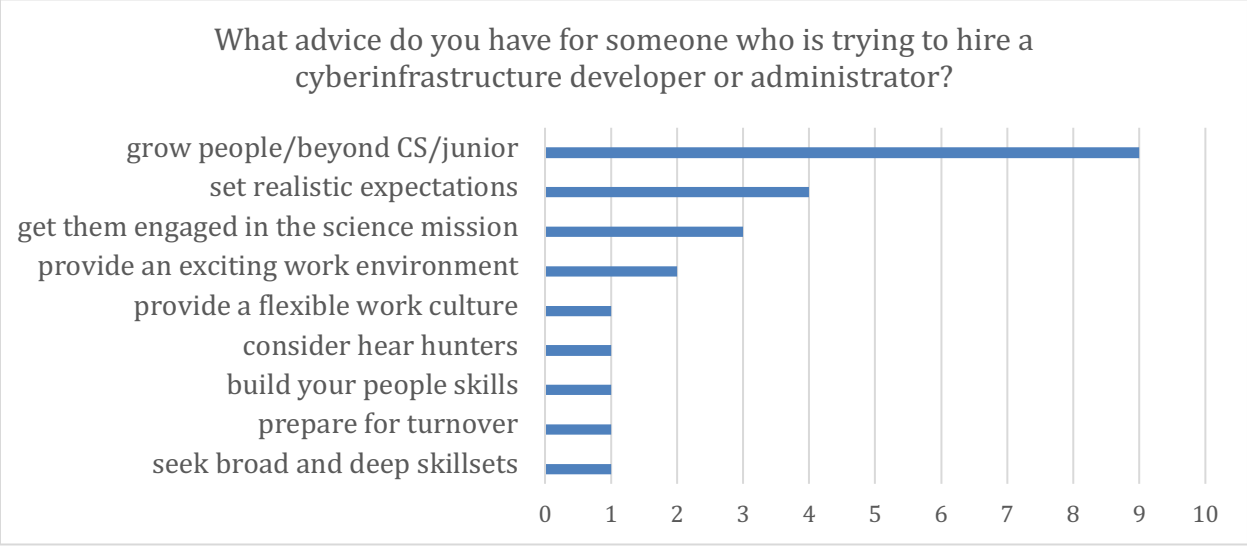
When asked how likely their facility/project would be willing to contribute resources to a shared workforce development effort, respondents did not feel strongly (mean=3.28, median=3 on the 5-point Likert scale). However, there were more highly positive responses than in the question about a shared service.

When asked what resources their facilities/projects would be willing to contribute to a shared workforce development effort, most respondents did not answer or indicated their facilities did not have relevant resources to share (50%). Those who specified concrete resources mentioned human resource/developers/personnel (9.38%) and expert/supervisor time (6.25%). There were also many unique responses offering both specialized resources and expertise.

Again, respondents were more willing to contribute expertise than resources to a shared workforce development effort, with a mean of 3.69 and median of 3.5 on the 5-point Likert scale. When asked what expertise they would be willing to share, the vast majority (59.38%) left the field blank or responded with "unsure" or "none." The two salient types of expertise mentioned by other respondents were broad experience in CI development and leadership (12.5%) and domain expertise and experience working with remote, shared CI (6.25%).

Finally, the respondents indicated that the most likely obstacles in implementing a shared workforce development effort would be similar to the obstacles perceived for implementing the shared services, most prominently the lack of funding, time, or personnel (47%). Other obstacles mentioned included lack of shared skills, divergent goals, competition with industry, scheduling, agreement to collaborate, and connectivity.

The workshop organizers also ask the participants if they had advice for hiring CI practitioners. Twenty-four participants answered this question in a free form paragraph. The graph below shows an analysis of the results.



The advice of training people by starting with junior people or people with a non-CS background was most frequently given. The responders also indicated that it is important to set a realistic expectation of what needs to be accomplished and what the personnel can expect. Engaging in the science mission and providing an exciting and flexible work environment were also indicated, although at a lower level, with 3, 2, and 1 response each. One participant also acknowledged that staff turnovers will happen and that you need to "hire personnel with strong organizational and documentation skills, and ability to think in multiple contexts."

Practitioner Survey

The workshop also conducted a very brief survey of CI practitioners associated with the workshop participants, with the aim to better understand knowledge sharing in the community. Seventy-eight responses were received with over 15 different LFs participating. Most of the responses came from CMS (14), the Academic Research Fleet (11), and NHERI (5) LFs. Six members of the Open Science Grid, which is a distributed CI serving LFs such as ATLAS, CMS, IceCube, and LIGO, also provided input.

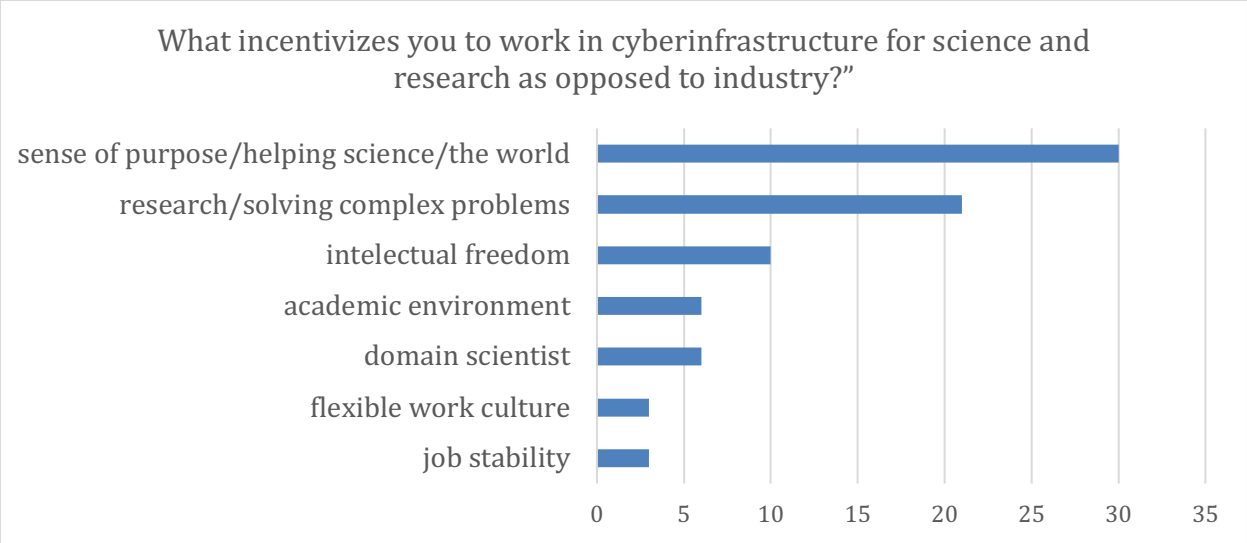
The workshop organizers also asked the CI practitioners, “Do you feel like you are part of a community? If so, what is it and how do you participate in your community?” Sixty-eight respondents answered yes, and 10 answered no—that they did not feel part of the community. Not surprisingly, members of the CMS LF, who have a large number of collaborators and who were the major responders to the survey felt a great sense of belonging to a greater community. An example of a response from that community included “*Yes. The USCMS program creates a strong community of computing professionals and maintains good internal communication.”*

Based on the responses to the second question, we aimed to distinguish between people belonging to a particular project, LF, rather than the broader CI community, for example, by attending conferences that are organized by outside entities. Although this is not a conclusive result because of the open-ended nature of the question, we would estimate that over 42% of respondents felt they belonged to a broader community of CI professionals.

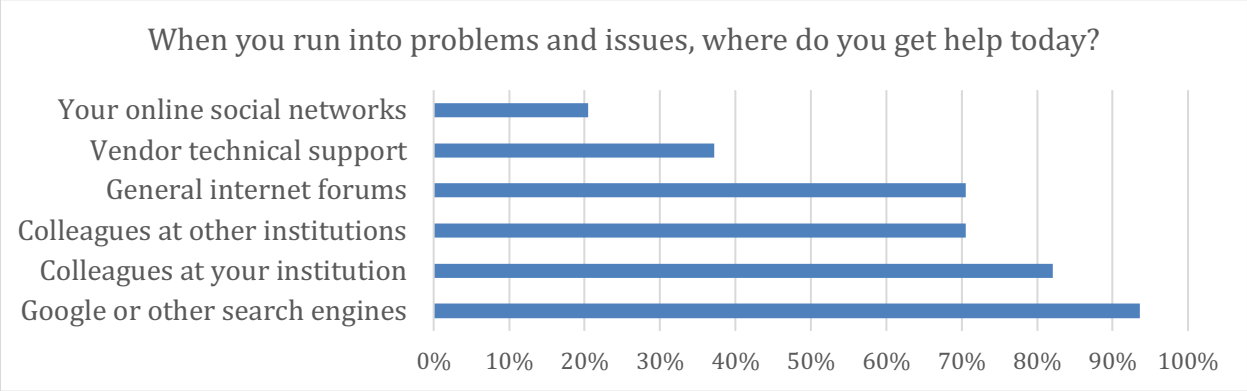
Since much of the workforce issues revolved around recruitment and retention, we wanted to explore why the CI practitioners pursue a career in academia or research laboratories. We asked, “What incentivizes you to work in cyberinfrastructure for science and research as opposed to industry?” The respondents were able to provide free text answers, and we categorized them as follows: having a sense of purpose, helping science and the world, the ability to conduct research and solve complex problems, having intellectual freedom to purse new ideas or contribute to open source projects, and working in an academic environment or having a flexible work culture or job stability. In some cases, CI practitioners were domain scientists that develop CI to meet their science objectives. An example of such a response was: “I am a professor of physics and care about the domain science deeply - not just providing general cyber-infrastructure.”

In some cases, the responders indicated multiple incentives for working in CI. The graph below summarizes the results. It is clear that having a sense of purpose and helping science was the foremost incentive for the responders (30). One respondent answered the question with “Belief that what we are doing helps the world.” Being able to conduct research and solve complex problems was the second most cite reason (21): “Simply put, it isn't boring. Science disciplines evolve and the needs and demands for infrastructure evolve with them. Matching technology to those needs is constantly changing and in need of "engineering" effort, so there is plenty to do!”

Although the sample of people surveyed was limited, the results do indicate that working in CI can provide a meaningful and attractive work environment. This finding can also potentially help recruit new personnel to careers in CI.

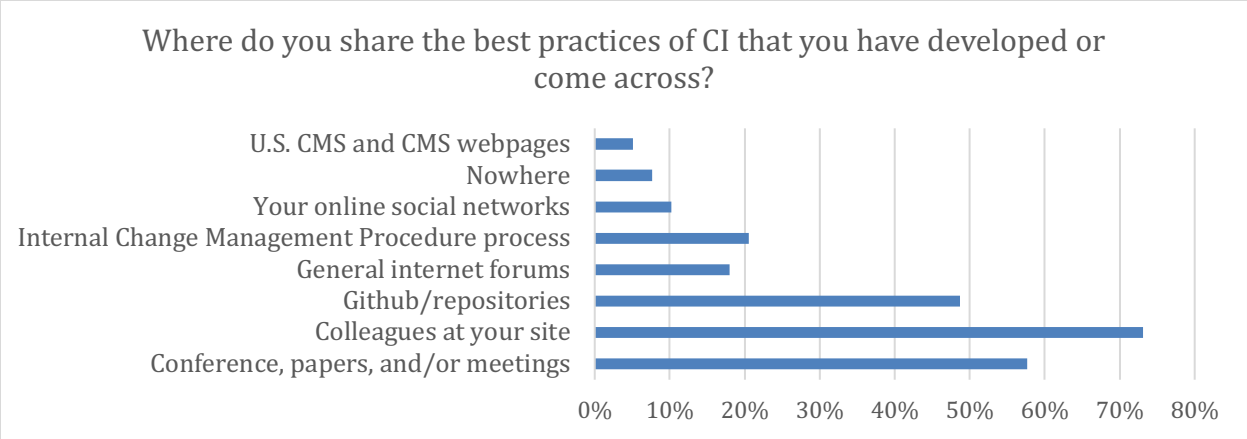


We asked CI practitioners about where they get help when they run into problems and issues. The respondents indicated a variety of sources (they were asked to check all the categories that applied):



Over 90% of respondents turn to Google Search for help with the CI issues, and over 70% make use of the general internet forums. On the other hand, the fact that 82% of respondents turn to CI colleagues at their institutions (and 70% to those at other institutions) suggests that CI-specific resources would be in high demand.

We find similar trends with where the CI practitioners share CI best practices they developed or come across:



Clearly, providing a centralized CI-related forum for both searching for and sharing the best CI practices would be very beneficial to the community.

Finally, the respondents were asked to specify their favorite CI resource. There were 32 different answers in total. The most common answers were Google and GitHub, which were both mentioned by 8 participants (10.5% of total participants). OSG was mentioned 5 times (6.6%). Stackoverflow and Campus Champions were both mentioned twice (2.6%). The remaining answers, very community specific, were all only mentioned once (1.3%), like CERN, CaRRC, RVTEC, and so on.

8. References:

[2017 LF CI report] 2017 NSF Large Facilities Cyberinfrastructure Workshop Report https://facilitiesci.github.io/assets/reports/facilitiesci-workshop-report-11-17.pdf

[2019 LF workshop] Pamplet about thr 2019 Large Facilities Workshop https://www.largefacilitiesworkshop.com/wp-content/uploads/2019/03/19LFWpamphlet.pdf

[arf] Academic Research Fleet https://www.unols.org/

[askc] Ask Cyberinfrastructure https://ask.cyberinfrastructure.org/

[carcc] Campus Research Computing Consortium https://carcc.org/

[casc] Coalition for Academic Scientific Computation https://casc.org

[chameleon] Chameleon https://www.chameleoncloud.org/

[cicoe-pilot] Cyberinfrastructure Center of Excellence Pilot https://cicoe-pilot.org/

[cip] Cyberinfrastructure Professional Intern Program

[earthcube] EarthCube https://earthcube.org/

[esip] Earth Science Information Partners https://www.esipfed.org/

[fair] Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg et al. "The FAIR Guiding Principles for scientific data management and stewardship." Scientific data 3 (2016).

[gemini] Gemini Observatory http://www.gemini.edu/

[htcondor] HTCondor https://research.cs.wisc.edu/htcondor/

[itil] Robert R. Moeller, Executive's Guide to IT Governance: Improving Systems Processes with Service Management, COBIT, and ITIL, 27 February 2013, ISBN:9781118138618, Online ISBN:9781118540176, DOI:10.1002/9781118540176.

[lccf] Leadership-Class Computing Facility https://lccf.tacc.utexas.edu/

[lf_list] https://www.nsf.gov/bfa/lfo/docs/major-facilities-list.pdf as of November 05, 2019

[lsst] The Large Synoptic Survey Telescope https://www.lsst.org/lsst

[mate] Marine Advanced Technology Education http://www.marinetech.org/internships/

[neon] The National Ecological Observatory Network https://www.neonscience.org/

[nheri] National Hazards Engineering Research Infrastructure https://www.designsafe-ci.org/facilities/experimental/

[noao] National Optical Astronomy Observatory https://www.noao.edu

[osg] Open Science Grid https://opensciencegrid.org/

[rcrv] Regional Class Research Vessel https://ceoas.oregonstate.edu/ships/rcrv/

[sgci] The Science Gateways Community Institute, https://sciencegateways.org/

[trustedci] Trusted CI, the NSF Cybersecurity Center of Excellence, https://trustedci.org/

[xsede] The Extreme Science and Engineering Discovery Environment (XSEDE) https://www.xsede.org/

Appendix A: Workshop Contributors

Steering Committee

The workshop was organized by a steering committee composed of CI experts from the LFs and selected CI projects:

- Brian Bockelman, Morgridge Institute and HTCondor project [htcondor]
- Adam Bolton, National Optical Astronomy Observatory [noao]
- Tom Cheatham, University of Utah and Campus Research Computing Consortium [carcc]
- Ewa Deelman (PI and Chair), University of Southern California and CI CoE Pilot [cicoe-pilot]
- Tom Gulbransen, Battelle, NEON [neon]
- Kate Keahey, Argonne National Laboratory, Chameleon [chameleon]
- Marina Kogan, University of Utah
- Dan Stanzione, Texas Advanced Computing Center, Leadership-Class Computing Facility [lccf]
- Daryl Swensen, Oregon State University, Regional Class Research Vessel [rcrv].

Discussion Leads

In addition to the steering committee, the following people led the breakout sessions and facilitated discussions:

- Ilya Baldin, RENCI, UNC-Chapel Hill
- Patrick Brady, University of Wisconsin-Milwaukee
- Rafael Ferreira da Silva, USC Information Sciences Institute
- Anirban Mandal, RENCI, UNC - Chapel Hill
- Jarek Nabrzyski, University of Notre Dame
- Susan Sons, Indiana University, CACR
- Alexander Szalay, Johns Hopkins
- Douglas Thain, University of Notre Dame
- John Towns, NCSA

Writing Contributors

The following people led the note taking during the meeting and contributed to this report:

- Ewa Deelman, USC Information Sciences Institute, Lead coordinating author
- Ilya Baldin, RENCI, UNC-Chapel Hill
- Laura Christopherson, RENCI, UNC-Chapel Hill
- Tom Gulbransen, Batelle
- Anirban Mandal, RENCI, UNC-Chapel Hill
- Angela Murillo, Indiana University-Purdue University Indianapolis
- Valerio Pascucci, University of Utah
- Steve Petruzza, University of Utah
- Mats Rynge, USC Information Sciences Institute
- Susan Sons, Indiana University, CACR
- Chaudhuri Surajit, Microsoft
- Charles Vardeman, University of Notre Dame
- Jane Wyngaard, University of Notre Dame

List of Attendees

First Name	Last Name	Organization/	Large Facility
Mark	**Abbot**	WHOI/	Ocean Observatories Initiative (OOI)
Rachel	**Adams**	UC Boulder/	Natural Hazards Center
Stuart	**Anderson**	Caltech	LIGO
Scot	**Arnold**	NSF	
Ilya	**Baldin**	RENCI, UNC - Chapel Hill	
Steve	**Barnet**	UW-Madison	IceCube Neutrino Observatory
Chaitan	**Baru**	NSF	
Karan	**Bhatia**	Google	
Cheryl Ann	**Blain**	Naval Research Laboratory	NHERI NCO
Brian	**Bockelman**	Morgridge Institute for Research	
Adam	**Bolton**	NOAO	National Optical Astronomy Observatory
Devin	**Bougie**	Cornell University	CHEXS
Patrick	**Brady**	University of Wisconsin-Milwaukee	LIGO
Jennifer	**Bridge**	University of Florida	NHERI
Joel	**Brock**	Cornell University	CHEXS
Sharon	**Broude Geva**	University of Michigan	
Peter	**Bryan**	Lehigh University	NHERI
Robert	**Chadduck**	NSF	
Vipin	**Chaudhary**	NSF	
Surajit	**Chaudhuri**	Microsoft	
Thomas	**Cheatham**	University of Utah	
Laura	**Christopherson**	RENCI, UNC - Chapel Hill	
Peter	**Couvares**	Caltech	LIGO
Chris	**Davis**	NSF	
Ewa	**Deelman**	USC Information Sciences Institute	
Mark	**Dufour**	NSF/DIAS/BFA	
Lee	**Ellett**	Scripps Institution of Oceanography, UCSD	U. S. Academic Research Fleet
Douglas	**Ertz**	UNAVCO	UNAVCO
Ken	**Feldman**	University of Washington	UNOLS SatNAG
Changda	**Feng**	Florida International University	WOW EF
Rafael	**Ferreira da Silva**	USC Information Sciences Institute	
Douglas	**Fils**	Consortium for Ocean Leadership (COL)	
Ian	**Foster**	ANL and University of Chicago	

Montona	**Futrell-Griggs**	NSF	
Rob	**Gardner**	University of Chicago	ATLAS
Philip	**Gates**	Texas A & M University	IODP
Erwin	**Gianchandani**	NSF	
Jeffrey	**Glatstein**	WHOI	Ocean Observatories Initiative (OOI)
Brian	**Glendenning**	NRAO	National Radio Astronomy Observatory
Tom	**Gulbransen**	Battelle Memorial Institute	NEON
David	**Halstead**	NRAO	National Radio Astronomy Observatory
Ben	**Hauger**	AURA, NOAO	National Optical Astronomy Observatory
John	**Haverlack**	University of Alaska	Academic Research Fleet
Pamela	**Hill**	NCAR	National Center for Atmospheric Research
Margaret	**Johnson**	NCSA/UIUC	LSST
JJ	**Kavelaars**	National Research Council of Canada	
Kate	**Keahey**	ANL	
Jeff	**Kern**	NRAO	National Radio Astronomy Observatory
Marina	**Kogan**	University of Utah	
Alexis	**Lewis**	NSF	
Sean	**Liddick**	NSCL / MSU	National Superconducting Cyclotron Laboratory
Miron	**Livny**	University of Wisconsin-Madison	
Eric	**Lyons**	University of Arizona	
Shadi	**Mamaghani**	NSF	
Anirban	**Mandal**	RENCI, UNC - Chapel Hill	
Thomas	**Marullo**	Lehigh University	NHERI Lehigh
William	**Miller**	NSF	
Angela	**Murillo**	Indiana University-Purdue University Indianapolis School of Informatics and Computing	
Jarek	**Nabrzyski**	University of Notre Dame	
Sanjay	**Padhi**	Amazon Web Services	
Manish	**Parashar**	NSF	
Valerio	**Pascucci**	University of Utah	
Joy	**Pauschke**	NSF/CMMI	
Chuck	**Pavloski**	Penn State Institute for CyberScience	
Steve	**Petruzza**	University of Utah	
Kevin	**Porter**	NSF	

Roland	**Roberts**	NSF	
Stefan	**Robila**	NSF	
Christopher	**Romsos**	Oregon State University, CEOAS	US Academic Research Fleet
Jim	**Rosser**	IODP JRSO	International Ocean Discovery Program · JOIDES Resolution Science Operator
Mats	**Rynge**	USC Information Sciences Institute	
Matt	**Schoettler**	UC Berkeley	NHERI SimCenter
Susan	**Sons**	Indiana University, CACR	
Dan	**Stanzione**	TACC, UT Austin	Leadership-Class Computing Facility
Laura	**Stolp**	WHOI	SatNAG/R2R
Alejandro	**Suarez**	OAC	
Daryl	**Swensen**	Oregon State University	Regional Class Research Vessel
Alexander	**Szalay**	Johns Hopkins	
Troy	**Tanner**	APL-UW	National Hazards Reconnaissance (RAPID)
Douglas	**Thain**	University of Notre Dame	
Christopher	**Thompson**	Purdue University	NHERI - Network Coordination Office
Kevin	**Thompson**	NSF	
Joanne	**Tornow**	NSF	
John	**Towns**	NCSA	
Charles	**Vardeman**	University of Notre Dame	
Ed	**Walker**	NSF	
Joseph	**Wartman**	University of Washington	National Hazards Reconnaissance (RAPID)
Von	**Welch**	CACR, Indiana University	
Tim	**Weston**	UC Boulder	
Dan	**Wilson**	UC Davis	NHERI CGM (Center for Geotechnical Modeling)
Frank	**Wuerthwein**	UCSD/SDSC	CMS
Jane	**Wyngaard**	University of Notre Dame	
Michael	**Zentner**	SDSC	

Appendix B: Agenda

Day 1 – Monday, September 16, 2019

07:30 – 08:30	Breakfast and Registration
08:30 – 08:35	Opening Remarks William Miller (NSF)
08:35 – 09:05	Building LF Cyberinfrastructure Communities to Advance the Endless Frontier Joanne Tornow (NSF)
09:05 – 09:40	Setting the Stage: 2017 CI workshop and the Cyberinfrastructure Center of Excellence Pilot Ewa Deelman and Tom Gulbransen
09:40 – 10:10	Guided Activity Kate Keahey and Rafael Ferreira da Silva
10:10 – 10:40	Break
10:40 – 12:00	Panel: State and Future of Cyberinfrastructure for Large Facilities Moderator: Dan Stanzione Panelists: Stuart Anderson, Margaret Johnson, Eric Lyons
12:00 – 13:00	Lunch Break
13:00 – 13:15	NSF/CISE Perspective Erwin Gianchandani (NSF)
13:15 – 13:45	Large Facilities Data Lifecycle Anirban Mandal
13:45 – 15:15	Lightning Talks Moderator: Mats Rynge Patrick R Brady: SCiMMA: Scalable Cyberinfrastructure to support Multimessenger Astrophysics Steve Petruzza: Interactive Access and Visualization of Large Scale Image Data Miron Livny: Sustaining Software Across a Growing Number of Facilities David Halstead: Next Generation Very Large Array: Communications Jeffrey Glatstein: Analysis of Alternatives Ian Foster: Cloud CI Services for Agile Facilities

	Chris Romsos: Software Defined Infrastructure in the Academic Research Fleet (ARF) Brian Glendenning: Scaling NRAO to the future Rob Gardner: New Approaches to Building and Operating Distributed Cyberinfrastructure for Large Facilities Douglas Fils: Schema.org and structured data for discovery Margaret Johnson: Enabling Multi-Instrument Pixel-Level Science with A High Throughput Computing, Data Access and Analysis Facility Kate Keahey: Chameleon: How to Build a Cloud++
15:15 - 15:30	Result Survey Overview and Setting up the Breakouts Ewa Deelman
15:30 – 16:00	Break
16:00 – 17:30	Parallel Breakouts: 1. What are the opportunities for collaboration between LFs and other Large CI projects? Leads: Ilya Baldin and Brian Bockelman 2. What are the CI challenges that need to be addressed in the next 5 years to support LFs science missions? Leads: Anirban Mandal and Daryl Swensen 3. What are the non-technical issues that influence CI development and how they can be collaboratively addressed? Leads: Susan Sons and Doug Thain
17:30 – 18:00	Breakout Summaries Top 3-5 findings and recommendations from each group
18:30 – 20:30	Reception

Day 2 – Tuesday, September 17, 2019

07:30 – 08:20	Breakfast
08:20 – 08:30	Setting the stage for Day 2 Ewa Deelman
08:30 – 10:00	Panel: Shared CI Services Opportunities and Challenges Moderator: Adam Bolton Panelists: Pamela Hill, JJ Kavelaars, Von Welch, and Mike Zentner
10:00 – 10:30	Break
10:30 – 12:00	Panel on Workforce Development and Retention Moderator: Tom Cheatham Panelists: Rachel Adams, Sharon Broude Geva, Jim Rosser, Frank Wuerthwein
12:00 – 13:00	Lunch Break

13:00 – 14:30	Parallel Breakouts: 1. Building a CI community: what are the impediments and opportunities? Leads: Patrick Brady and Marina Kogan 2. Enhancing the CI workforce: what are the challenges and solutions? Leads: Jarek Nabrzyski and John Towns 3. CI for science, where does the LF CI end and the user CI begin? Leads: Tom Gulbransen and Alex Szalay
14:30 – 15:00	Breakout Summaries Top 3-5 findings and recommendations from each group
15:00 – 15:15	Wrap-up

Appendix C: Cyberinfrastructure Calling Cards

The following CI Calling Cards have been collected from the attendees prior to the workshop.