NSF LARGE FACILITIES

CYBERINFRASTRUCTURE






Report from the NSF Large Facilities Cyberinfrastructure Workshop

Alexandria, VA, USA, September 6-7, 2017

NSF Large Facilities Cyberinfrastructure Workshop[1]

Alexandria, VA, USA, September 6-7, 2017

http://facilitiesci.org/

Table of Contents

[1] The workshop was support by the National Science Foundation through grant number ACI 1742969

Executive Summary

Summary of the Workshop

The 2017 NSF Large Facilities Cyberinfrastructure Workshop was held at the Westin Alexandria in Alexandria, VA on September 6-7, 2017. The workshop was organized by a steering committee composed of experts from the facilities as well as from the CI community. Workshop attendance was by invitation, and consisted of 82 attendees including representatives from the NSF large facilities, the cyberinfrastructure (CI) community, and the National Science Foundation (NSF). The overarching goal of the workshop was to enable direct and synergistic interactions among the NSF large facilities and the CI communities to jointly address the CI needs as well as the sustainability of the CI of existing and future large facilities.

The workshop was preceded by a pre-workshop survey and the collection of white papers from the facilities. The 26 responses to the survey questionnaire as well as the 22 white papers were shared with the attendees prior to the workshop and made available to the general public though the workshop website. The workshop program was developed on the basis of the survey responses and was primarily composed of 4 panels focused on key crosscutting issues followed by breakout discussions on each panel topic.

The white papers and survey responses provide a wealth of information about facilities' CI and its operation. A summary of the survey responses is presented in this report, and the responses and the white papers are included as appendices. An analysis of the white papers to identify crosscutting requirements and challenges, architectural patterns, effective practices, and opportunities for sharing interoperability is an important key next step.

Overall, the workshop provided a unique forum for interaction and frank discussions between the facilities and CI communities on important issues related to the facilities' CI, and was viewed as constructive by all participants. Workshop details as well as related material are available at http://facilitiesci.org/.

Summary of Panels and Breakouts

1. *Integration, interoperability and reuse of CI solutions, practice*: Facilities typically address CI challenges independently of each other and develop custom solutions in an uncoordinated manner, missing the opportunity to leverage existing solutions and knowledge. Facilities can benefit from a trusted forum that can facilitate discussions, collect and disseminate information about addressing technical challenges, and provide information and potential evaluation of existing CI solutions. Such a forum could form a bridge between existing facilities and also help provide expertise for when new facilities start up, or support continuity when the responsibility for facility operations is transferred to a new group.
2. *Workforce development, and education and outreach*: Facilities are facing significant workforce development, education, and outreach challenges, including promoting and

maintaining a highly skilled CI workforce, workforce diversification, and continuous learning, while encountering poor mission alignment to host institution HR policies. However, they are independently working to address these challenges. Fostering community learning based on independent programmatic successes and lessons learned via increased intra-facilities communications has the potential to form effective, network-wide workforce strategies.

3. *CI models, challenges, best practices*: Sharing best CI practices, e.g., core tools, systems, is valuable, and such best practices exist across the facilities. A common location of knowledge, system descriptions, and use cases was seen as highly desirable to the community. In addition, the community suggested a topic-specific conference focused on CI best practices.
4. *Sustaining Facilities CI / Developing a community*: There needs to be a long-term commitment to the continuity and sustainability of core CI services and end-to-end processes, as well as personnel and knowledge. The processes and budgetary structures underlying facilities do not support refactoring, evolution, and sharing of CI, or its interoperability, with other facilities. An external entity that provides expertise and knowledge services across facilities can be a critical resource to make CI more effective and sustainable. Developing a facilities' CI community can be extremely beneficial; however, there are currently no mechanisms or incentives to support the development of such a community.

Key Findings and Recommended Actions

The key findings and corresponding recommended actions summarized below are synthesized by the steering committee from the pre-workshop survey responses, white papers, and discussions and feedback from the community before, during, and after the workshop, including the pre- and post-workshop surveys and workshop panels and breakouts. There was also an overall view among the participants that while this workshop was effective and resulted in useful discussions and insights, a longer term sustained activity that combines crosscutting as well as topic-focused discussions is needed to fully address the challenges and opportunities related to large facilities' CI.

Key Findings

- 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.).
- There is a lack of effective mechanisms and funding structures to support interactions and sharing among facilities regarding their CI. There is also a lack of a facilities' CI community that can collectively address CI sustainability and help provide continuity between existing and future facilities.
- There is a need for, and a current lack of, easily accessible information about current CI technologies, solutions, practices, and experiences.

- There is a critical lack of a focused entity that could facilitate interactions and sharing across facilities. A model such as that used by the NSF-funded Center for Trustworthy Scientific Cyberinfrastructure (CTSC) was explicitly and repeatedly noted as an effective model that should be explored to address this gap.
- The constantly changing technology and CI landscape highlights the tradeoffs between longer stability and incorporating new and potentially more effective/efficient solutions. A crosscutting approach for addressing these tradeoffs is missing, as are mechanisms and funding for evolving/refactoring facilities' CI.
- Workforce development, training, retention, career paths, and diversity are major crosscutting challenges that the community shares. They may be best addressed coherently across all facilities through a coordinated approach.

Recommended Actions

- Foster the creation of a facilities' CI community and establish mechanisms and resources to enable the community to interact, collaborate, and share.
- Support the creation of a curated portal and knowledge base to enable the discovery and sharing of CI-related challenges, technical solutions, innovations, best practices, personnel needs, etc., across facilities and beyond.
- Establish a center of excellence (following a model similar to the NSF-funded Center for Trustworthy Scientific Cyberinfrastructure, CTSC) as a resource providing expertise in CI technologies and best practices related to large-scale facilities as they conceptualize, start up, and operate.
- Establish structures and resources that bridge the facilities and that can strategically address workforce development, training, retention, career paths, and diversity, as well as the overall career paths for CI-related personnel.

1. Introduction

Overview and Goals

Cyberinfrastructure is a critical component of NSF facilities. As the CI grows in scale and complexity, it is essential for the facilities and CI communities to collectively explore how to most effectively provide and sustain their essential components and services to meet current and future needs. The 2017 NSF Large Facilities Cyberinfrastructure Workshop was held on September 06 and 07 in Alexandria, VA, with the overarching objective of enabling direct and synergistic interactions among the NSF large facilities and the cyberinfrastructure communities to jointly address the CI needs and sustainability of existing and future large facilities. A key goal was to develop a common understanding of the current and evolving requirements, architectures, and best practices; enabling technologies; operation practices and experiences; and issues and gaps.

Specific goals of the workshop included:

- Understand best practices of current CI architecture and operations at the large facilities.
- Identify common requirements and solutions as well as CI elements that can be shared across facilities.
- Enable CI developers to most effectively target CI needs and the gaps of large facilities.
- Explore opportunities for interoperability between the large facilities and the science they enable.
- Develop guidelines, mechanisms, and processes that can assist future large facilities in constructing and sustaining their CI.
- Explore mechanisms and forums for evolving and sustaining the conversation and activities initiated at the workshop.
- Generate recommendations that can serve as inputs to current and future NSF CI-related programs.

Workshop Organization and Attendees

Steering Committee

The workshop steering committee was responsible for organizing the workshop, including conducting the pre-workshop activities, developing the workshop agenda, coordinating the workshop activities, and producing the workshop report. The committee was composed of leading experts from the facilities as well as from the CI community. The committee members were as follows (bios are included in Appendix B):

- Stuart Anderson, California Institute of Technology and LIGO
- Ewa Deelman, Information Sciences Institute (ISI), University of Southern California
- Manish Parashar (PI and Chair), Rutgers University and OOI
- Valerio Pascucci, University of Utah
- Donald Petravick, NCSA, University of Illinois, Urbana-Champaign and LSST

- Ellen M. Rathje, University of Texas at Austin and NHERI

Workshop Attendees

Workshop attendance was by invitation. The workshop attendees included representatives from the NSF large facilities and the CI community. Representatives from other agencies (e.g., Department of Energy facilities), industry (e.g., Microsoft), as well as international facilities (e.g., Ocean Networks Canada) were also invited. The workshop was attended by 86 participants (out of 98 registrations). Travel support was provided to two early career researchers, Rafael Ferreira da Silva from USC Information Sciences Institute, and Eliu A. Huerta from the University of Illinois at Urbana-Champaign. The final attendee list is included as an appendix.

Workshop Structure and Activities

Pre-workshop Activities

The pre-workshop survey focused on gathering information about the facilities' CI deployments and associated issues to (1) provide material ahead of time to the workshop attendees, and (2) provide information to the steering committee to develop the most effective structure and agenda for the workshop. Specifically, the pre-workshop survey included two components:

Questionnaire: The facilities were requested to answer the set of questions listed below. The goal of the brief questionnaire was to provide information about the facilities' CI to enable the steering committee to plan the workshop and develop its agenda.

White papers: Attendees were requested to upload a short (up to 2 pages in length) white paper. White papers were expected to provide an overview of the facility and its CI; the design, deployment, and operation of the CI; and key CI capability/service and/or best practice, as appropriate.

We received 26 responses to the questionnaire and 22 white papers, representing a large fraction of the invited facilities. All responses to the pre-workshop survey, i.e., the white papers and the answers to the questionnaires, were shared with the workshop attendees prior to the workshop, and are attached to this report as appendices.

The white papers and survey responses provide a wealth of information about facilities' CI and its operation. A summary of the survey responses is presented in this report in Section 2, and the responses as well as the white papers are included as appendices. An analysis of the white papers to identify crosscutting requirements and challenges, architectural patterns, effective practices, and opportunities for sharing interoperability is a key next step.

Workshop Website: A website dedicated to the workshop was created at http://facilitiesci.org, and is used to communicate details about the workshop to the attendees, including the objectives of the workshop, the results of the pre-workshop survey (white papers, responses to

the questionnaire), the workshop structure and agenda, and workshop logistics. This final workshop report will also be posted on the workshop website.

Timeline for Pre-Workshop Activities: Pre-workshop activities occurred during late spring and early summer 2017. White papers and surveys were due by the end of June. Once these were received, the steering committee used them to develop the workshop agenda. The committee also created summaries of the responses and presented them at the workshop.

The Workshop

The 2017 NSF Large Facilities Cyberinfrastructure Workshop was held at the Westin Alexandria in Alexandria, VA on September 06 and 07, 2017. Workshop attendance was by invitation, and consisted of 82 attendees composed of representatives from the large facilities, CI community, and NSF. The workshop program was derived from the responses to the pre-workshop survey, and was primarily composed of 4 panels focused on key crosscutting issues followed by breakout discussions on each panel topic. Additionally, there was an opening summary and discussion of the survey results, and invited presentations by Irene Qualters from NSF and James Marsteller from the Center for Trustworthy Scientific Cyberinfrastructure. Overall, the workshop provided a unique forum for interaction and frank discussions between the facilities and CI communities on important issues related to facilities' CI, and was appreciated by all participants.

Post-Workshop Activities / Deliverables

Post-Workshop Survey: A brief post-workshop survey was conducted to get feedback from the attendees on the structure and content of the workshop, as well as their views on follow-up activities. The responses to the survey are included as an appendix to this report.

Workshop Report: The steering committee has produced this report as an outcome of the workshop, which captures the results of the preparatory activities as well as the discussions, findings, and recommended actions from the workshop. This report will be posted on the workshop website and will be publicly available.

2. Summary of Pre-workshop Survey

Pre-workshop Survey Questionnaire

1. What significant components of the CI were developed in-house? Are these components available to others to reuse?
2. What external CI capabilities and services and/or externally developed tools (if any) does the facility use and who provides them? How were these tools identified and what criteria were used to select the tools?

3. List up to 3 of your most used and most challenging CI components with a 1-sentence explanation for each. What aspects of the facility CI and its operation would you like to share as best practices?
4. What aspects of the facility CI and its operation do you see as challenges or gaps? Are there "CI lessons learned" that you would like to share or see discussed at the workshop?
5. What do see you as key risks in facility CI (e.g., dependency on external resources such as compute, data, expertise, and/or services)? Are there mitigation steps that you would like to share or see discussed at the workshop?
6. What CI-related workforce development activities does your facility engage in?
7. What do you see as your key new CI requirements and challenges in the next 5-10 years?
8. Do you have any other suggestions for the workshop?

Summary of Survey Responses

What significant components of the CI were developed in-house? Are these components available to others to reuse?

A majority of the CI components are developed in-house because of the need for tailored solutions to deal with a particular environment and facility needs, such as, for example, specific sensor data capture, distribution and replication, instrument control, etc. Such in-house development results in a complete suite of services and tools for a particular community. While the software is often made available open source to others, the reuse of such software suites is unclear. There is also some software development to support business processes such as procurement.

What external CI capabilities and services and/or externally developed tools (if any) does the facility use and who provides them? How were these tools identified and what criteria were used to select the tools?

There is significant reuse of systems software across the facilities, such as, for example, systems for software development (e.g., Confluence, Jenkins, Github), web development (e.g., Apache servers, Django, various DBMS systems), software distribution (e.g., Docker and Singularity containers), and authentication (e.g., Shibboleth, GSI, LDAP). NSF-funded CI software such as Globus (especially GridFTP), HTCondor, Pegasus, and THREDDS are also used by some of the facilities. There is also wide use of CI platforms such Open Science Grid and XSEDE. Several facilities leverage the capabilities and knowledge provided by the Center for Trustworthy Scientific Cyberinfrastructure (CTSC). Finally, some facilities are leveraging cloud technologies for data management.

In terms of how these tools are identified and what criteria were used to select the tools, facilities seem to rely on their IT staff, technical teams, and governing committees. Some

facilities have used more structured requirement gathering and evaluation of various existing software solutions before making a selection.

List up to 3 of your most used and most challenging CI components with a 1-sentence explanation for each. What aspects of the facility CI and its operation would you like to share as best practices?

The respondents reported that most challenging CI components are related to data, networking, and computing. In the case of data, challenges include addressing rapidly growing multi-petabyte datasets, ingesting data from distributed sources, and providing efficient and effective access to data. In the case of networking, challenges include addressing reliability and high-bandwidth requirements as well as dealing with international scales in both collecting and redistributing data. Computing challenges include managing large and diverse workflows, deploying and using services such as Jupyter, and using elasticity for HTC components.

Best practices highlighted in the responses include using systems engineering to manage CI lifecycle and interfaces, and integrating redundancy in the CI design to provide high availability.

What aspects of the facility CI and its operation do you see as challenges or gaps? Are there "CI lessons learned" that you would like to share or see discussed at the workshop?

Challenges and gaps identified in the survey responses spanned budgetary, recruiting and retention, technology and operation, and security.

Budgetary challenges mentioned in the responses include handling growing data and user communities with shrinking budgets, and the growing costs of keeping CI (hardware and software) adequately provisioned and up to date. The responses also noted the struggle between the costs of using commercial services and the uncertainties associated with using services provided by academics.

Technology and operations challenges include handling evolving requirements and technology and the integration of new components, dealing with increasing data rates and scaling CI capacity and performance, migration to cloud platforms, operating (and coordinating) widely distributed CI and addressing associated bandwidth requirements, balancing new technologies and ideas with CI stability, handling growing computing requirements at the core as well as the edges, and the pervasive need for integrating Integration of "legacy" software/solutions.

The security challenge highlighted by respondents is the ongoing struggle of minimizing security risks while maintaining open access and supporting international users.

"Lessons learned" highlighted in the responses include the benefits of implementing industry best practices for CI deployment and operation, ensuring the traceability of CI features to

requirements and business needs, building redundancy (storage, networking, VM clusters, connectivity) into the CI architecture, leveraging third-party services when possible, and establishing structures for communication and interaction across different teams and groups.

What do see you as key risks in facility CI? Are there mitigation steps that you would like to share or see discussed at the workshop?

The key risks mentioned in the survey responses include risks related to funding, infrastructure and technology, and workforce.

Funding risks mentioned are those associated with increasing costs of CI operations and management, evolution, data sustainability, personnel, etc.

Infrastructure and technology risks highlighted include the growing scales and complexity and CI and associated management and operation challenges, handling technology disruptions, avoiding technology and/or vendor lock-in, ensuring access to required computing resources (HPC and clouds), ensuring connectivity, especially in the case of wide-area or low bandwidth CI, and handling the unreliability of sensors and other instrumentation.

Workforce risks mentioned center around workforce development, recruiting and retention, and handling the loss of key personnel and associated knowledge.

Integration and interoperability risks highlighted in the responses are the challenges of sharing data as well as knowledge, expertise, and infrastructure, while scalability risks are focused on the growing scale and diversity of the user community and associated budgetary and technical challenges.

Security risks noted include the focus on the growing cybersecurity-threat landscape and the challenge of securing CI while maintaining usability and access and ensuring user productivity.

Mitigation strategies noted included communicating with funding agencies; sharing CI components, services, and practices; developing data lifecycle management systems; implementing redundancies; and leveraging enterprise technologies, services, and practices (e.g., for management).

What CI-related workforce development activities does your facility engage in?

Workforce development and retention was noted as one of the top priorities and risks as well-trained personnel are needed to perform tasks effectively and they are also harder to retain. The survey responses also indicated a high variability of approaches, ranging from simply “allowing” personnel to participate in training activities to supporting personnel participation in workshops and other meetings, arranging coordinated visits from external speakers and professionals,

organizing annual activities such as summer schools and annual conferences, and organizing combinations of commercial and research/academic activities.

Training and workforce development activities at facilities include both professional development/training for technical staff and training for the user community and students. The former include technical training on technologies and best practices, certification programs, monthly webinars and internal training sessions, establishing mentoring relationships among junior and senior staff, encouraging technical staff to attend conferences/workshops, and organizing joint meetings with other projects/programs. The latter include workshops for students and faculty, involving students in operation and research, setting up summer programs, and hosting summer interns.

Workforce development and training challenges noted in the responses include the observation that facilities can involve a workforce with a wide variety of seniority and expertise (e.g., students, postdocs, professors, scientists, developers, managers, etc.), and it is hard to "unify" training across the spectrum. The responses also highlighted diversity challenges, and the fact that there is little or no systematic internal training/mentoring and no budget allocation for it.

Statistics derived from the responses indicate that training and workforce development activities at facilities largely focus on technical staff and include technology training at workshops and conferences and through online mechanisms. The statistics also highlighted that training for managers is largely nonexistent.

What do you see as your key new CI requirements and challenges in the next 5-10 years?

Anticipated CI requirements and challenges in the near future (5-10 years) noted in the survey responses spanned data, computing and networking, software, operations and maintenance, integration and interoperability, workforce development and training, and community engagement.

Data requirements/challenges include exploding data volumes/rates and the associated need for scaling CI capacity and performance; leveraging new techniques such as machine learning; ensuring high-speed, real-time delivery of data; incorporating novel data delivery mechanisms; and addressing longer term data archiving.

Computing and networking challenges mentioned in the responses include the increasing need for capacity and capabilities, handling technology disruptions, identifying the role of cloud services, ensuring high-bandwidth wide-area network links, and handling the growing complexity of sensors and instruments.

Software-related challenges include the long-term stability of software and the reproducibility of scientific results.

Operations and maintenance challenges highlighted in the responses are the need for configuration, management tools, developing SLAs, and the growing cybersecurity concerns.

Integration and interoperability challenges noted by respondents include the need for the integration of facilities and data across facilities and the complexity of inter-federation agreements, particularly involving international partners.

Highlighted responses regarding workforce development and training include the training and retention of professionals and the training of “teachers”.

Finally, community engagement challenges include handling increasing user demand and supporting a growing community.

Do you have any other suggestions for the workshop?

- What is NSF’s vision / role in coordinating and/or providing (facilities) CI through grant funding?
- What are best practices for using cloud services for large-scale science data storage and access?
- What are the mechanisms, structures, and incentives for sharing and interoperability?
- How can we sustain the community and conversations beyond the workshop?

3. Panel on Integration, Interoperability, and Reuse of CI Solutions, Practices

Goals and Structure of the Panel

As noted in the introduction, cyberinfrastructure is a critical component of NSF facilities. It is used to direct instruments, collect data from sensors, transfer data across wide-area networks, process large amounts of data, and conduct simulations, among others. Cyberinfrastructure includes services, software, and hardware that make communications, data management, and computing possible. Each NSF facility is faced with the deployment and potential development of cyberinfrastructure components. During the process of planning, development, deployment, and operation, a significant amount of knowledge is gained, so the question is how we transfer this knowledge. As the responses to the survey indicate, there is a considerable amount of effort spent by the facilities on developing in-house CI solutions. Since the development of CI can be costly in time and monetary terms, it is important to understand how this CI can be developed in a way that integrates and interoperates with existing solutions. A related issue is also how this CI can be made available in a way that is reused by others so that the investments made by NSF can be leveraged across the facilities and even more broadly by the NSF community. Oftentimes developers consider putting the source code and documentation on Github, which is a good step toward reuse but not enough. Leveraging CI investments is

important but so is the need to sustain these investments over time. CI is a living entity, which needs to be maintained, adapted, and potentially expanded as time goes on and the CI ecosystem and user and facility needs change. Thus, it is necessary to develop methods and procedures for facilities and CI communities to provide and sustain essential CI components and services to meet current and future needs of facilities and the NSF community.

The panelists included members of facilities and CI: Aaron Anderson (National Center for Atmospheric Research), Gonzalo Merino (IceCube Neutrino Observatory), Michael Zentner (NanoHub), and Kate Keahey (Chameleon Cloud). The panel was charged with answering the following questions:

- How do your projects discover and evaluate available solutions?
- How do your projects deal with the changing availability of CI?
- Can increased awareness/reuse of CI solutions increase interoperability across facilities?
- Can community efforts in integration, interoperability, and sustainability lead to well-defined interfaces that facilitate access to and incorporation of new technologies?
- What are the most critical CI gaps that you would like to see addressed?

Observations by the Panel

How do your projects discover and evaluate available solutions?

The panelists used a number of different venues to discover available solutions, ranging from one-on-one interactions to engagement with community bodies to community venues. These venues included partner interactions (including outside the NSF community), vendor interactions, bodies of expertise (Science Gateways Institute, the Center for Trustworthy Scientific Cyberinfrastructure), and conference/workshops (CI focus). Some panelists also stated that it was important to be connected to a range of different communities: high-throughput computing, high-performance computing, and various domain sciences that are relevant to the facility. In terms of evaluation of the CI, there was no single model. Based on requirement gathering, and analysis, the evaluation of solutions is based on a number of criteria:

- source code availability
- size of the supporting team and the user community
- costs and risk analysis, which looks at potential problems as the facilities evolve.

The discovery and evaluation processes cannot be done just at one point in time but rather continuously as new user needs emerge and new CI solutions become available.

How do your projects deal with the changing availability of CI?

All panelists agreed that changes in CI are constant and rapid, and thus there is a constant risk and cost for the facilities to adapt and there is an impact on the user community. Panelists discussed the need to use formal agile methods to be able address a changing CI environment. There is also a tension between staying with existing solutions versus evaluating and adopting new, sometimes less stable, CI. In some cases, one can reach out to user support services provided by national cyberinfrastructures such as the Open Science Grid and XSEDE. In order

to be able to make such decisions, the process of community and stakeholder engagement needs to be sustained throughout the lifetime of the facility.

The Open Science Grid (OSG) was described as an infrastructure that enables processing among a diverse number of both university-based facilities as well as XSEDE and DOE computing centers. The central concept of the Open Science Grid is to provide operational "plumbing" that makes the ensemble of computing and storage of over 100 US sites available for use to a large number of projects. The OSG allows collaborators at multiple institutions to build a unified distributed collaboration computing in an organized way. The OSG allows unused cycles at a site to be used by outside collaborations when that site specifically consents.

Can increased awareness/reuse of CI solutions increase interoperability across facilities?
The panelists felt that increased reuse has the potential to increase interoperability, but that it needs to be supported by the business model, which can exploit potential synergy in generating results, and take into account the limitations on funds and thus foster pooling of resources. Interoperability can also be fostered by programmatic mandates. There are also challenges to interoperability, for example, the complexity of developing interoperable solutions and the costs of potentially extra development. Sometimes the solutions may not be directly applicable and may limit the functionality of the overall solution.

Can community efforts in integration, interoperability, and sustainability lead to well-defined interfaces that facilitate access to and incorporation of new technologies?
There are some success stories in integrating various community solutions. For example, a number of projects utilize HTCondor for their job management solutions and leverage the Globus Data Transfer Services for data movement. The ESNet Data Transfer Node architecture has also been adopted by some projects. However, these are very low-level capabilities, and there are opportunities to leverage more complex solutions.

What are the most critical CI gaps that you would like to see addressed?
The panel illuminated a number of gaps and opportunities. At the fundamental level, they noted the need for information sharing, including:

- Discoverability and availability of services and software solutions.
- Data management capabilities: a number of facilities struggle with data management. It is hard to handle the complexity and volume of data, and thus sharing solutions, formats, and ontologies may be beneficial.
- There can also be sharing of best practices in terms of data curation, long data preservation and sharing, and reproducibility of scientific results.

There is also a need to **keep up with trends** as the computing landscape becomes more complex (heterogeneous systems), including developing software/services/system evaluation methods and potentially sharing the results of the evaluations with an understanding that various facilities have different needs and evaluation criteria.

Given the dynamic computing landscape and the complexity of systems and applications, **automation techniques** are needed to manage the data lifecycle (including computing) on behalf of the end user.

Sustainability of CI tools and services was identified as an important component; however, there was a distinction made between sustaining a particular software or service versus sustaining a particular capability, which may have different implementations.

There was a point made that technologies such as **machine learning techniques** that are used in the business domain to select products for consumers can also be used to enhance our scientific productivity, creativity, and collaboration.

Technological advances need to be supported by sound computer science, and thus some of the gaps identified dealt with the issue of **robust science**, which supports repeatability and reproducibility and the management of the provenance of the data sets collected. Robust science also includes the rigorous treatment of experimental testbeds, such as Chameleon, to include testbed versioning, appliances, experiment management, and replay, among others. Publication of software as well as data was also viewed as important.

Not all gaps that were identified were technical; some were **cultural**. There is a need to incentivize developers and service providers to make their CI accessible and discoverable. Funding agencies can also incentivize projects to reuse existing solutions when developing new capabilities. There is also often a communication gap between domain scientists and computer scientists and IT providers. Bridging this gap is important because interdisciplinary partnerships are viewed as keys to success.

Discussion

A number of points were raised during the discussion. Some touched upon the issue of trust when using someone else's software or services: Can you trust the software to perform as expected? Can you trust that the software/service will be supported and issues addressed when problems arise? Will the software/service be available in the long term? This concern brought up the different timescales at which the facilities and the "outside" CI components are funded. The former often spans decades whereas CI software and services are funded in 5-year intervals in the best case.

All facilities need to make decisions about CI adoption and weigh the risks and the rewards. It is also clear that there are commonalities in the facilities' needs: authentication, account management, data movement, and computation management, among others. There was a discussion regarding creating a forum for information and experience sharing, potentially along the lines of CTSC, which was viewed as a great example of a community resource. The appealing part of CTSC is that it provides solutions without promoting a particular piece of software. A common forum could also help with the discovery and evaluation of software and services and assist with the development of a "facilities blueprint" that could help facilities get off

the ground and develop in an efficient and sustainable way, by leveraging existing CI solutions where available and developing new CI where gaps exist.

Communication between facilities has been recognized as an important component. Communication should occur at all levels of the organizations (PIs, developer, system administrators, etc.).

Information sharing can address not only current issues, but also the challenges that lie ahead. For example, HPC systems are changing and becoming more complex and heterogeneous. However, the computational requirements of applications include both HPC and HTC. Commercial clouds offer services that are becoming more attractive for science. However, understanding the cloud cost model is not trivial.

Data management is also a critical aspect of facilities. Some facilities deal with remote and challenging conditions, for example when collecting data at sea or at the South Pole. Many facilities deal with data management across the wide-area networks (especially across administrative domains) and in terms of issues of processing and access to large data sets. Data access for end users is also becoming a bottleneck-- going to the website and downloading the data is not sustainable.

Conclusion

The panel and subsequent discussions raised a number of challenges that facilities are addressing in various and often uncoordinated fashions. In response to these challenges, facilities are often developing their own custom solutions, missing the opportunity to leverage existing solutions and knowledge. Facilities, and more broadly CI projects funded by NSF, and even end-users can benefit from a trusted forum that can facilitate discussions, collect and disseminate information about addressing technical challenges (data and computation management, etc.), and provide information and potential evaluation of existing CI solutions. This forum could also undertake the support and maintenance of services and software that are deemed critical to the facilities. This forum could not only form a bridge between existing facilities but also help provide expertise for when new facilities start up, or support continuity when the responsibility for facility operations is moved to new group.

4. Panel to Workforce Development, Education, and Outreach

The NSF facility workforce is a critical aspect of the facility and, along with many other aspects, such as cyber infrastructure components, requires active investment and development. The NSF facilities and associated CI workforce are highly specialized and represent a significant investment by the NSF facility community. A facility's workforce holds a myriad of responsibilities, including: sustains current operations of the facility, interacts and supports facility users, selects and advances the facility cyber infrastructure, and performs or supports

both education and outreach activities. Additionally, the facility CI workforce provides the ability for facilities to: (i) evaluate and adopt rapidly evolving CI; (ii) incorporate knowledge gained from direct experience and experience of others in best practices, including commercial and open-source solutions; (iii) provide feedback to the community of NSF infrastructure projects; (iv) ensure that facilities are optimally relevant to the science communities served; and (v) maintain contacts and foster new interactions amongst NSF facilities and their workforces.

The theme of this panel and discussion evolved naturally along the 3 highly interrelated issues of development, outreach, and education.

Development. The development of a technical CI workforce garnered enthusiastic discussion during the workshop. Of the many topics addressed, retention, development, and diversity were the most robustly discussed, with many challenges being acknowledged, which, in turn, led to a conversation about several nascent strategies in early or informal practice at isolated facilities. Significantly, participants in both the panel and breakout session firmly pointed out that workforce development in CI efforts differed for scientific, technical, and administrative staff.

Education. Many of the NSF facilities are closely associated with universities, requiring a look at training as both skill development for senior, permanent personnel, and traditional education, which is largely motivated by the presence of students and postdocs both in the facility's workforce and among its users. Thus, training and educating in these instances must also address preparing junior staff and others for employment outside the facility or institution, likely in industry or academics.

Outreach. Discussion during the workshop on outreach specific to a workforce was relatively limited. Several NSF facilities, such as the Uindata effort, have outreach intrinsically embedded in their mission, but outreach specifically focusing on a workforce is not well addressed. Several discussions cited the potential for increased multidisciplinary workforce outreach as a possible mechanism for growing technical, ethnic, race, and gender diversity into the NSF facilities workforce.

Goals and Structure of the Panel

The members of the panel were drawn from NSF facilities: Albert Lazzarini (LIGO), Mohan Ramamurthy (Unidata Program Center), Aaron Andersen (National Center for Atmospheric Research), and Ellen Rathje, (Natural Hazards Engineering Research Infrastructure).

Questions to prime the panel discussion were:

- What are workforce development-related challenges and best practices, e.g., recruitment, training, incentives, reward structures, etc.?
- How can we better facilitate awareness and sharing of solutions and best practices across different and disparate communities?
- How can training and outreach become key mechanisms for concurrent improvement of workforce effectiveness and the ability to adopt best tools and understand future needs?

- How similar or different are the training needs of our community with respect to industry?
- Can industry practices and resources be leveraged?
- Is there a proper balance between technical and managerial training?
- Does the funding model properly emphasize workforce development?
- Does the workforce training properly address the full career path?

Observations by the Panel

All NSF facilities have programs related to development of the facility's workforce, education, and outreach. These programs face very similar challenges, and in many cases there exist informal or nascent programs addressing these shared challenges. However, the responses to workforce challenges are largely, if not exclusively, done in an isolated fashion. With similar challenges there is significant potential to develop shared best practices.

In general, world-class large facilities are able to attract the efforts and attention of world-class scientists and postdoctoral fellows. However, there are a variety of challenges in maintaining a consistent and experienced CI workforce, some of which are shared whereas others are specific to the various NSF facilities:

Recruiting and Retention: A common thread regarding workforce at NSF facilities is that the challenges associated with recruiting and retention differ significantly among scientific, technical, and administrative workforces. The world-class scientific missions at world-class NSF facilities create a natural attraction for a world-class scientific workforce, where membership in the facility's workforce is a natural part of a scientific member's career path. For the technical and administrative workforce, this natural alignment of career interests is not as well defined. As such, salaries enabled by NSF awards are not generally competitive for the CI skillsets that are also in high demand within private industry. Adding to the difficulty of competing against higher salary offerings is the complexity and limited flexibility of HR polices within academic institutions. Additionally, the career paths within facilities for the technical and administrative workforce are not as evident relative to the scientific workforce. These factors are challenges in both the recruiting and retention of the technical and administrative workforce.

Oftentimes, NSF facilities are in desirable work locations, which often aids in recruitment of the workforce, but with desirable work locations come disadvantages to both recruitment and retention. Certainly a high cost of living is a striking disadvantage exacerbated by the relatively low pay and inflexible pay scales of many NSF facilities.

Adding to the concerns of salary competitiveness is the funding nature of NSF-supported missions. Staff development and retention is seldom directly supported within NSF budgeting parameters, and grant budget cycles lack the assurance of long-term employment for staff members. Thus, given the finite duration of the awards given to facilities, staff are constantly aware of the need to remain marketable to outside entities; above and beyond the need to stay current in new technologies. This apparent lack of long-term employment for CI staff also

creates a limit to building community and staff commitment and is another challenging factor in workforce development and retention.

Further, human resource (HR) practices within awardee host organizations, such as universities, may not be optimal for HR management of a facility, although the nature of the practices needing improvement varies greatly from institute to institute. Examples of HR practices noted as creating challenges for maintaining a CI workforce include job titles being restricted to those applicable to university IT, and inflexible HR rules regarding salaried versus hourly employment (e.g., rules that cause system administrators to become hourly employees). The inflexible HR practices/positions, along with the relatively short timeframes of facility budgets, often diminish the sense of commitment and professionalism of CI professionals and offer relatively limited career ladders. In fact, there may exist a disconnect between the facility leadership and the understanding of staff development within host institutions. In the workforce development breakout session, a straw poll was taken and the great majority of attendees were not able to say that they understood the staff development processes at their institutions.

Each facility is responding to these various challenges. Several approaches are either being practiced or envisioned. Certainly, acknowledging the facilities' workforce has several constituents, including scientific, technical, IT/CI, and administrative, with each constituency presenting unique challenges, was seen as universally important. For instance, maintaining a current technical knowledge is a marked challenge for the IT/CI workforce, whereas retention is a marked challenge in the administrative workforce.

Several current practices and several ideas for fostering recruitment and retention were offered. Using existing models and programs such as SOARS (Significant Opportunities in Atmospheric Research and Science), REU (Research Experiences for Undergraduates), and internships was seen as a common practice for bringing new staff members into the facilities. Also, a leading practice for recruiting and retaining scientists and staff is to use the scientific mission of the facility or CI group as a recruiting tool. The opportunity to work on and support work on an important scientific topic often overcomes wage disparity and leads to a high degree of commitment. Another compensating factor that accompanies CI employment within NSF facilities is the more academic and flexible work environment, relative to industry, offered within the facilities. More directly, to overcome lagging salaries, off-cycle bonuses are used to create higher effective pay. It was pointed out that not every institution has this capability and many budgets do not support this strategy.

It was also suggested that some components of the workforce could be recruited from industry; specifically, those workforce members looking to move from the intense and restrictive nature of industry to the more flexible, academic environment of the facilities and the facilities' host organizations.

A specific example of an event that accomplishes both a technical end and also offers community building and professional development is that of the LIGO site's information security peer reviews. These peer reviews provide a structured and constructive opportunity for

information security staff from various facilities to interact, exchange information, and further their individual development. This facilitation of information exchange helps the facilities and also offers a sense of community and learning among the reviewers.

A noted need that resonated with the workshop attendees was that of creating career paths, beyond typical NSF budget periods, for facility staff, possibly even career paths that include moving among NSF facilities and CI providers.

Several discussions cited the potential for an increased multidisciplinary workforce outreach as a possible mechanism for growing technical, ethnic, race, and gender diversity in the NSF, essentially broadening the recruitment net. Stepping stones for this broadening include reworking job descriptions focusing on keywords used and technical requirements. Oftentimes, the job postings are very narrow, which artificially limits the applicant pool.

Discussion

A primary concern regarding workforce revolves around retention and the potential for the loss of experience and knowledge from the NSF facilities. This concern is especially important to the sustainment of those facilities with highly specialized missions and CI components. A primary retention challenge is the inherent limitations of facilities to satisfy the career needs of staff.

Limitations mentioned include:

- Short-term and inflexible budgets that provide for wage inflation, but not for wage increases or time lengths correspondingly needed to provide a stable career ladder for facilities' staff members.
- No specific budgeting mechanism for staff career and professional development.
- HR practices at host institutions the operate facilities do not align well with facilities and career development of staff members.
- Staff members who have limited awareness of information interchange with other NSF programs and projects having similar CI needs. These interactions would be a source of career-enhancing job satisfaction.

Secondary to retention as a discussion point was the idea of the benefit of professional development to the facility. In addition to helping increase retention, professional development strategies also aid in:

- Staff members gaining knowledge of technical/working best practices across other facilities and industry.
- Encouraging a high degree of commitment to the facility.
- Offering a strategy that may compensate for relatively lower wages compared to industry.

Many facilities have programming aimed at recruitment, retention, and to some degree professional development. These strategies are generally done in an isolated matter and are

more informal and *ad hoc* in nature, but offer a potential foundation for formal, shared programs to be developed across the NSF facilities. Examples of this programming include:

- Leveraging the world-class scientific missions of the facility to garner a workforce naturally committed to the facility's particular mission.
- Recruiting industry members based on the academic institution working environment flexibility.
- Outside, peer audits of facility components (e.g., IT security), which enable staff to engage in rich information exchange with peer groups.
- Encouraging and enabling attendance at conferences, workshops, and other professional development events.
- Using bonuses as a compensating factor to offset relatively low salaries.

In addition, workshop attendees also discussed workforce diversity. It was proposed that the group should look at rewriting position statements to use a more general language, which would attract a broader workforce applicant in many regards.

Conclusion

Many challenges exist in promoting and maintaining a highly skilled CI workforce. Challenges such as workforce diversification and continuous learning in a rapidly changing environment are shared by the entire tech community: industry, academics, and facilities alike. Other challenges such as restricted budgets, poor mission alignment of host institution HR policies, and servicing highly specialized facility needs are unique to facilities. It is reasonably evident that all facilities are independently working to address workforce challenges. Making available community learning from independent programmatic successes and lessons learned via increased intra-facilities communications has the potential to form effective, network-wide workforce strategies.

5. Cyberinfrastructure models, challenges, and best practices

Goals and Structure of the Panel

In general, the CI systems associated with large facilities have been developed and maintained individually without any coordination and using different architectures and models. This panel (and breakout session) explored the different models used to architect, design, construct, and maintain cyberinfrastructure components for large facilities. The panelists included experts from the facilities (Ivan Rodero, Ocean Observatories Initiative, Tom Gulbransen, NEON) and cyberinfrastructure (Frank Wuerthwein, Open Science Grid) communities. Specific questions addressed by the panel were:

- What are the existing models / best practices for architecting/acquiring and operating CI resources/services (storage, compute expertise, software)?
- What role can academic and commercial service offerings (e.g., cloud services) play?

- What are the advantages/disadvantages of developing shared CI resource / service models ?
- How can CI take advantage of rapidly evolving technologies and enterprise solutions?
- What best practices for updating/upgrading CI are already in place?
- Do facilities attempt to "lean forward" or does continuity/robustness trump any CI risk (e.g., balance of continuity versus forward-looking)?
- What partnerships have worked? Failed (e.g., other Federally Funded Research and Development Centers (FFRDC), academic, industry, etc.)?

Observations by the Panel

The panelists provided insightful responses to the questions above, which led to an active discussion, with the advantage of a variety of backgrounds and perspectives.

Activities involving data collection, processing, storage, and access are critical across all the facilities and constitute a core of competencies that need to be shared. Unfortunately, the facility responsibility often ends once the facility provides data access to researchers. Certainly, different researchers/communities have different needs, requirements, and skill levels regarding the use of data and corresponding CI tools, and so it can be difficult to adequately serve communities with heterogeneous needs using shared systems. However, there is a need for development and documentation of best practices and for training and engaging each research community to use the data provided with the best technologies available.

Some facilities have taken advantage of shared services and CI, with varying levels of success. Advantages include achieving critical mass to scale services, developing larger communities with common needs, and arriving at consensus over protocols. At the same time, it becomes much more challenging to schedule access to the available resources and to develop a uniform quality of services for the varying needs of the community.

In general, there is no single platform that can be easily used to exchange ideas, experiences, and opinions about the CI implemented by the different large facilities. A recurring theme throughout the panel and breakout session was that it would be useful to create a common forum for discussing and sharing CI issues across all facilities. An example of useful shared information is the architecture diagrams for various facilities' CI. To facilitate comparison, it is important for these diagrams to be developed using a common conceptual framework and consistent level of detail.

Discussion

Best practices are highly dependent on the circumstance of the respective user. In general, the workshop attendees felt that core system-level capabilities and services offer the highest potential for benefits from sharing and adopting best practices. For instance, sites could adopt one another's solutions for authentication and single sign-on across sites.

Often, the best practice is to use tools from industry. The use of industry tools comes with a specific set of challenges, which includes planned or unplanned obsolescence and the fact that best practices continually evolve, sometimes outside a company's specific product. An example of an industry-supplied tool is LIGO's use of the StorageTek/Sun Microsystems tape systems. Oracle acquired the StorageTek product line through an acquisition of Sun Microsystems. Since the tape libraries are not a central part of Oracle's corporate plan, the StorageTek tape libraries no longer offer support – a case of obsolescence through corporate acquisitions. At the same time, GEMINI has found that the support of a commercial storage solution is much more robust as compared to an open-source solution. However, this level of support comes with a price, specifically, a high rate of reinvestment in licenses and maintenance contracts. However, there was general consensus that industry best practices should inform CI construction, operation, and management, and should be leveraged where appropriate.

Of course, a common question when discussing CI best practices is how cloud services fit the facility CI model. Recently, some groups, such as CADC, have found the sweet spot where Amazon Web Services (AWS) can effectively meet the project's needs. Others, such as the CMS community, have found cloud services to still be prohibitively expensive. The lesson learned is that cloud services, whether for compute needs or data storage/access needs, are appropriate in some situations but not in others. Additionally, the decision to use cloud services incorporates both technical and budgetary questions (i.e., is it easier or more beneficial, from a budgetary perspective, to purchase equipment rather than invest in services, who pays for power, how much effort would it take to move software to a cloud model, etc.). A facility's funding structure and the balance between construction and operation funds also impacts this decision. Certainly, the potential of a systematic approach resolving many redundancies across facilities through a single, cloud-based solution is appealing, but requires solving a myriad of technical and other challenges. The topic is complex enough and the consensus was that it should be the focus of a dedicated separate discussion within the community.

The overall discussion generated many examples of highly specific mission or data use-based practices; for instance, use cases where the user wants to watch 10s from a video in an archive, which requires specific data annotation/indexing tools. The high variation in data scales across the facilities was also noted, which could necessitate moving analysis tools into a shared computational system processing the data, for example, using the cloud to perform the analysis. Many projects also have very specific needs, such as moving data from a telescope in Maui in Hawaii to a data center in Colorado, which is a challenge at 20TB per day and using cloud services is not in the near-future plans of the project. The relatively large percentage of such highly specific CI needs is perhaps an indication that early coordination and sharing activities should focus on on best CI practices at the core.

Conclusion

Several core tools or systems can be seen as best CI practices across the facilities. In some cases, especially at the core levels, these best practices have the potential of being shared systems. As the purpose of the tool or system nears the specific scientific mission or user, the

ability to establish best practices or share tools across facilities is less apparent. This thought is clear in the discussion of where and how the cloud can help provide CI for facilities. The community has a strong sense that sharing best practices is highly valuable when CI is considered. A common location of knowledge, system descriptions, and use cases was seen as highly desirable to the community. In addition, the community suggested a topic-specific conference focused on CI best practices.

6. Sustaining Facilities Cyberinfrastructure (CI) / Developing a Community

Goals and Structure of the Panel

Recognizing that CI has become a critical component of NSF facilities and is growing in terms of its scale and complexity, this sequence of the panel and breakout sessions explored approaches and challenges in sustaining essential CI components and services to meet current and future needs. The panel further explored the importance of a community of CI personnel in sustaining and evolving facilities' CI and how such a community could be developed and nurtured. The panelists included experts from the facilities (Tim Ahern, IRIS, and Dan Stanzione, NHERI) and cyberinfrastructure (Miron Livny, OSG, and John Towns, XSEDE) communities. Specific questions addressed by the panel were:

- What are the dimensions of CI sustainability (storage, software, expertise)?
- How are sustainability decisions made (e.g., what to sustain and what not to sustain)?
- What are the main barriers to sustaining CI solutions, both within facilities and externally?
- What are possible models for CI sustainability?
- Can sharing, reuse, and interoperability provide pathways to sustainability? Can community/market models play a role?
- How can we create and sustain a community to facilitate sharing and sustainability?

Observations by the Panel

The panelists provided insightful responses to the questions above, which led to an active discussion.

On the question of dimensions of CI sustainability, the panelists emphasized that the focus should be on the sustainability of services (compute, data, networking, expertise, etc.) and end-to-end processes, rather than specific tools, technologies, or providers. Furthermore, sustaining personnel and knowledge is an important aspect of overall CI sustainability, given the reality of changing technologies and evolving requirements. The importance of data and its attributes, including velocity, volume, variety, and veracity, was also highlighted, as were associated domain-specific needs, which often lead to unique (domain-specific) solutions. The panel largely

agreed that availability of resources and expertise is central to sustainability decisions, i.e., what aspects of CI to sustain or not sustain, and these decisions are often made at the agency level.

The panel noted several barriers to the sustainability of facilities' CI, including the existing legacy and domain-specific solutions that are often monolithic, challenges in retaining key CI personnel and the resulting churn and loss of knowledge, evolution of technology and the lack of backward compatibility and/or continued vendor support, and the lack for effective models for creating scalable shared CI. Establishing trust between CI users and providers was also noted as a challenge. The panel also pointed to cultural barriers to sustainability, such as the view that every CI is unique (snowflake problem) and the "not built here" phenomenon. An overarching concern noted was the fact that funding tends to favor innovation rather than sustainability, and that there is no funding model of CI sustainability. Related to this concern was the perception by some in the community that sustainability is more a service than research, which was pointed out as a barrier. Finally, the panel noted that sustaining CI often does not include refactoring or evolving it, for example, to achieve great functionality and/or efficiencies.

Key models of sustainability discussed by the panel are based on sharing (resources, operating costs, personnel) and leveraging other investments, which could lead to improved efficiencies and reuse. While such sharing/leveraging may not always be possible, for example, in the case of domain/facility-specific data processing, it should be explored for common capabilities (e.g., data lifecycle management, data delivery mechanisms, computing) and services (e.g., training, user support, expertise). Furthermore, while considering models based on consolidated service providers across facilities, the panel noted the need to maintain competitiveness. The panel also suggested that commercial cloud services also provide potential solutions as well but leveraging these services has been challenging. Overall, the panel noted that long-term sustainability requires a model where common capabilities/services become "utilities" and there is a commitment as a community (funding agencies, universities, researchers, etc.) to maintain the continuity of these services.

Discussion

The discussion during the panel and the breakout largely followed the questions that framed the panel. It reemphasized people as a key dimension of sustainability, but noted the need to sustain the entire CI ecosystem. It particularly noted models such as the Center for Trustworthy Scientific Cyberinfrastructure (CTSC), which provides expertise and services across facilities as a potential approach toward making CI more sustainable. The discussion also highlighted the need for ensuring the flow of new ideas/innovations (e.g., machine learning techniques, streaming data delivery mechanisms, etc.) into any sustained solutions, as well as the importance of creating community knowledge bases of effective practices and experiences and community discussion forums.

In terms of making sustainability decisions, the panelists noted that market models, which are essentially reactive, should be complemented with strategic planning, as well as effective metrics to support sustainability decisions. Similar costs issues were highlighted as key

concerns in build versus reuse decisions, as was the observation that there are costs even when reusing or outsourcing. The lack of vehicles for exploring longer term sustainability, for example, beyond the lifetime of the facility was highlighted. The discussion cautioned against monocultures, and that it was important to focus on best practices rather than standards.

The discussion at the breakout noted that CI sustainability is not the same as data archiving/preservation, and that there are government-funded entities that are focused on data archiving.

The discussions on barriers to sustainability highlighted the lack of budgetary structures. A similar observation was made about the evolution, sharing, and interoperability of CI components -- these aspects are not explicitly supported by the existing funding structures, which are focused on operating the CI as-constructed.

The discussion on model for sustainability focused on growing the community of users and reuse. However, it was noted that there are few resources and/or incentives for communities to come together to address CI and its sustainability in a collective way. Funding to support collaborative efforts across disciplines and facilities was discussed as an option. The discussion also highlighted that CI sustainability requires multidisciplinary efforts, and significant cultural barriers across communities often need to be overcome. Finally, it was noted that the sustainability and continuity of core and crosscutting services should be considered beyond the facilities, at a national level.

Conclusions

A summary of the highlights from the panel and breakouts session on sustainability is as follows:

- There needs to be a long-term commitment to the continuity and sustainability of core CI services and end-to-end processes that can be leveraged across facilities, rather than specific tools, technologies, or providers.
- Sustaining personnel and knowledge are critical aspects of the overall CI ecosystem, and their sustainability must be addressed across facilities.
- The processes and budgetary structures underlying facilities do not support refactoring, evolution, or sharing of CI, or its interoperability with CI from other facilities.
- An external entity that provides expertise and knowledge services across facilities can be a critical resource to make CI more effective and sustainable.
- While developing a facilities' CI community can be extremely beneficial in increasing the effectiveness and efficiency of CI and CI personnel, there are currently no mechanisms or incentives to support the development of such a community.

7. Findings and Recommended Actions

The 2017 NSF Large Facilities Cyberinfrastructure Workshop was attended by a total of 86 participants (out of 98 registrations) representing the facilities and CI communities, industry, and NSF. Overall, the workshop provided an effective forum of interaction and frank discussions between the facilities and CI communities on important issues related to facilities' CI, and was appreciated by all participants. The key findings and corresponding recommended actions summarized below are synthesized by the steering committee from the pre-workshop survey responses, white papers, and discussions and feedback from the community before, during, and after the workshop, including the pre- and post-workshop surveys and workshop panels and breakouts. There was also an overall view among the participants that while this workshop was effective and resulted in useful discussions and insights, a longer term sustained set of activities that combines crosscutting as well as topic-focused discussions is needed to fully address the challenges and opportunities related to large facilities' CI.

Findings

- There is a crosscutting desire across the facilities and CI communities to engage with each other, and to have a clear understanding of the advantages and impact of such an engagement. There are examples of existing, very successful interactions between facilities and CI, such as LIGO, IceCube, NHERI, etc. There are also examples of facilities successfully leveraging solutions developed by the CI community, such as HTCondor, GridFTP, Pegasus, etc. The clear benefits of interaction and sharing of architectures, practices, and experiences among the facilities (and with other entities such as FFRDCs) were also highlighted.
- There is a lack of effective mechanisms and funding structures to support interactions and sharing among facilities regarding their CI. There is also a lack of a facilities' CI community that can collectively address CI sustainability and help provide continuity between existing and future facilities.
- There is lack of, and need for, an online resource focused on CI technologies, practices, and experiences that can be kept up to date and easily accessed by the community. There is an overarching need for ensuring continuity and sustaining core services/utilities, such as data preservation, computing, etc. It is important that this continuity occurs across facilities and independent of providers and technologies, the insurance of which will also address the perceived risks with leveraging/reusing CI components and solutions.
- There is a critical lack of a focused entity that could facilitate interactions and sharing across facilities. A model such as that used by the NSF-funded Center for Trustworthy Scientific Cyberinfrastructure (CTSC) was explicitly and repeatedly noted as an effective model that should be explored to address this gap.
- There were open and frank discussions about gaps within facilities' CI, as well as challenges to interactions, collaborations, and sharing of practices, experiences, and solutions among the facilities and between the facilities and CI communities, which are resulting in lost opportunities and inefficiencies. Key among these challenges are limited

resources, a lack of structures and mechanisms, and a lack of incentives for such interactions, collaborations, and sharing. It was also noted that cultural barriers limit such interactions as well as sharing.

- The constantly changing technology and CI landscape highlights the tradeoffs between longer stability and incorporating new and potentially more effective/efficient solutions. A crosscutting approach is required for addressing these tradeoffs. Related is the lack of clear mechanisms and funding for evolving/refactoring facilities' CI within the funding structure. The importance of better coupling cycles of innovation across the facilities and CI communities was noted, especially for addressing the increasing complexity of CI requirements and solutions.
- The importance (as well as lack) of well-defined metrics and mechanisms for evaluating and comparing CI solutions was highlighted.
- The lack of appropriate career paths and reward structures for CI personnel at academic institution and recruiting and retention are crosscutting challenges that were highlighted.
- Workforce development, training, and outreach were noted as critical crosscutting needs across both communities. The lack of management training was particularly noted. The need for diversity in the facilities' CI personnel was also identified as a critical crosscutting concern across both communities. The lack of understanding of how best to address (and fund the addressing of) these issues was noted.
- The benefits of fostering and nurturing a community of CI personnel and its impact on, for example, perception of stability and realization of full career paths, as well as the current lack of such a community, were noted.
- The need for additional interactions with industry and the exploration of industry services was noted. In particular, understanding the role of commercial cloud services in terms of cost/benefits (e.g., identifying goldilocks zones) was highlighted.

Recommended Actions

- Foster the creation of a facilities' CI community and establish mechanisms and resources to enable the community to interact, collaborate, and share. Leverage such a community to facilitate career advancement and professional development for CI personnel. Develop culture and reward structures that encourage collaboration and sharing of technology.
- Support the creation of a curated portal and knowledge base to enable the discovery and sharing of CI-related challenges, technical solutions, architectural patterns, innovations, best practices, personnel needs, etc., across facilities and beyond. Such a resource could also collect experiences of technology adoption (e.g., commercial cloud technologies/services).
- Establish a center of excellence (following a model similar to the NSF-funded Center for Trustworthy Scientific Cyberinfrastructure, CTSC), as a resource providing expertise in CI technologies and best practice related to large-scale facilities as they conceptualize, start up, and operate. Such a center would, for example, facilitate communication and sharing of experiences, practices, CI elements, and risks, as well as help facilities explore the benefits and impact of technology changes/disruptions, etc.

- Establish structures and resources to strategically address workforce development, training, retention, career paths, and diversity, as well as the overall career paths for CI-related personnel at facilities. Collect best practices for recruiting, training, workforce development, retention, etc.
- Develop shared (standardized) metrics as well as methodologies for evaluating and computing software and other CI elements.
- Explore collaborations and synergies with facilities funded by other agencies, as well as with industry.

Appendix A. Agenda

Day 1 – Wednesday, September 06, 2017

07:30 – 08:30 Registration/Breakfast

08:30 – 09:00 Welcome / Goals of the Workshop
Irene Qualters and William Miller, NSF / Steering Committee

09:00 – 10:00 Setting the stage (Review of survey responses + Discussion)
Steering Committee

10:00 – 10:30 Break

10:30 – 12:00 Panel 1 (4 panelists; short talks + q/a session)
Focus: Integration, interoperability and reuse of CI solutions, practices

- How do your projects discover and evaluate available solutions?
- How do your projects deal with changing availability of CI?
- Can increased awareness/reuse of CI solutions increase interoperability across facilities?
- Can community efforts in integration, interoperability and sustainability lead to well defined interfaces that facilitate access to and incorporation of new technologies?
- What are the most critical CI gaps that you would like to be addressed?

Moderator: Ewa Deelman, USC/ISI
Panelists: Aaron Anderson (NCAR), Gonzalo Merino (IceCube), Mike Zentner (NanoHub), Kate Keahey (Chameleon)

12:00 – 01:00 Lunch

01:00 – 02:30 Panel 2 (4 panelists; short talks + q/a session)
Focus: Workforce development, and education and outreach

- What are workforce development related challenges and best practices, e.g., recruitment, training, incentives, reward structures, etc.?
- How can we better facilitate awareness and sharing of solutions and best practices across different and disparate communities?
- How can training and outreach become key mechanisms for concurrent improvement of workforce effectiveness and ability to adopt best tools and understand future needs
- How similar or different are the training needs of our community with respect to industry? Can industry practices and resources be leveraged?

Moderators: Valerio Pascucci, University of Utah,
Donald Petravick, NCSA/UIUC

Panelists: Ellen Rathje (NHERI), Aaron Andersen (NCAR), Albert Lazzarini (LIGO), Mohan Ramamurthy (UCAR)

02:30 – 03:00 Break

03:00 – 04:30 Breakouts for Panel 1 (Room: Banneker; Lead: E. Deelman & K. Keahey)
Breakouts for Panel 2 (Room: Bell; Lead: V. Pascucci and D. Petravick)

04:30 – 05:00 Report out from breakouts

05:00 – 05:30 Planning for Day 2

06:00 – 09:00 Reception

Day 2 – Thursday, September 07, 2017

07:30 – 08:15 Breakfast

08:15 – 08:30 Summary of Day 1 / Goals of the Day 2

08:30 – 09:00 Invited Talk 1: NSF/OAC Update, Irene Qualters, CISE/OAC

09:00 – 09:30 Invited Talk 2: Summary of Cybersecurity Summit – J. Marsteller, CTSC

09:30 – 10:00 Break

10:00 – 11:30 Panel 3

Focus: CI models, challenges, best practices

- What are the existing models / best practices for architecting/acquiring and operating CI resources/services (storage, compute expertise, software) ?
- What role can academic and commercial service offerings (e.g., cloud services) play ?
- What are there advantages/disadvantages of developing shared CI resource / service models ?
- How CI take advantage rapidly evolving technologies and enterprise solutions ?

Moderator: Ellen Rathje, NHERI/University of Texas at Austin

Panelists: Kerstin Lehnert (IEDA), Ivan Rodero (OOI), Frank Wuerthwein (LHC/OSG), Tom Gulbransen (NEON)

11:30 – 12:30 Lunch

12:30 – 02:00 Panel 4

Focus: Sustaining Facilities CI / Developing a community

- What are the dimensions of CI (storage, compute expertise, software) sustainability?
- How are decisions about sustainability (e.g., what to sustain and what not to sustain) made?
- What are the barriers (if any) towards sustaining CI solutions, both within facilities and externally?
- What are possible models for CI sustainability?
- Can sharing, reuse, interoperability provide pathways to sustainability? Can community/market models play a role?
- How can we create and sustain a community to facilitate sharing and sustainability ?

	Moderator: Victoria Stodden (UIUC), Manish Parashar, Rutgers University **Panelists:** Miron Livny (OSG), Tim Ahern (IRIS), Dan Stanzione/Ellen Rathje (NHERI), John Towns (XSEDE)
02:00 – 02:30	Break
02:30 – 04:00	Breakouts for Panel 3 (Room: Banneker; Lead: E. Rathje) Breakouts for Panel 4 (Room: Bell; Lead: M. Parashar)
04:00 – 04:30	Report out from breakouts
04:30 – 05:00	Workshop closing

Appendix B. Steering Committee Bios

Stuart Anderson (PhD, California Institute of Technology) has 25 years of experience working on computational intensive astronomy and physics experiments. He is currently a staff scientist at Caltech where he leads the computing program for LIGO. His experience includes building radio astronomy instrumentation, collecting and analyzing large time domain data sets in a small research group to discover relativistic binary pulsar systems using high-performance computing (HPC) techniques, working in a large science collaboration to discover ultra-relativistic binary black hole systems using high-throughput computing (HTC) techniques, and managing the data analysis computing and archival systems for LIGO.

Ewa Deelman is a research professor at the USC Computer Science Department and a research director at the USC Information Sciences Institute (ISI). Dr. Deelman's research interests include the design and exploration of collaborative, distributed scientific environments, with particular emphasis on automation of scientific workflow and management of computing resources, as well as the management of scientific data. Her work involves close collaboration with researchers from a wide spectrum of disciplines. At ISI she leads the Science Automation Technologies group that is responsible for the development of the Pegasus Workflow Management software. In 2007, Dr. Deelman edited a book: “Workflows in e-Science: Scientific Workflows for Grids”, published by Springer. She is also the founder of the annual Workshop on Workflows in Support of Large-Scale Science, which is held in conjunction with the Supercomputing conference. In 1997 Dr. Deelman received her PhD in Computer Science from the Rensselaer Polytechnic Institute.

Manish Parashar is Distinguished Professor of Computer Science at Rutgers University. He is also the founding director of the Rutgers Discovery Informatics Institute (RDI2). His research interests are in the broad areas of parallel and distributed computing and computational and data-enabled science and engineering. Manish is founding chair of the IEEE Technical Consortium on High Performance Computing (TCHPC), EiC (elect) of the IEEE Transactions on Parallel and Distributed Computing (TPDS), serves on the editorial boards and organizing committees of a large number of journals and international conferences and workshops, and has deployed several software systems that are widely used. He has received a number of awards for his research and leadership. Manish is a fellow of AAAS, a fellow of the IEEE/IEEE Computer Society, and an ACM Distinguished Scientist. For more information please visit http://parashar.rutgers.edu/.

Valerio Pascucci is the John R. Park Inaugural Chair of the University of Utah and the founding director of the Center for Extreme Data Management Analysis and Visualization (CEDMAV) at the University of Utah. Valerio is also a faculty of the Scientific Computing and Imaging Institute, a professor in the School of Computing, University of Utah, and a laboratory fellow of PNNL. Before joining the University of Utah, Valerio was the Data Analysis Group Leader of the Center for Applied Scientific Computing at Lawrence Livermore National Laboratory. Valerio's research interests include big data management and analytics, progressive multi-resolution techniques in scientific visualization, discrete topology, geometric

compression, computer graphics, computational geometry, geometric programming, and solid modeling. Valerio is the coauthor of more than 200 refereed journal and conference papers and is an associate editor of the IEEE Transactions on Visualization and Computer Graphics.

Donald Petravick (B.S. University of Illinois at Chicago) has 32 years of experience working in support of survey astronomy and high energy physics. He currently heads the Astronomy Core Services Department at the National Center for Supercomputing Applications, where his duties include principal investigator for the Dark Energy Survey Data Management System, Operations Architect for the LSST Data Management Subsystem, and local principal investigator for activities at NCSA supporting the LSST project. His experience includes computing facilities, large data storage frameworks, international wide-area networks, information security planning, high-throughput computing, software development, and management for both real-time and offline systems. He also spent a year as a detailee to the Department of Energy Office of High Energy Physics, where he obtained a level of understanding about agency program management.

Ellen M. Rathje is the Warren S. Bellows Centennial Professor in the Department of Civil, Architectural, and Environmental Engineering at the University of Texas at Austin (UT), and senior research scientist at the UT Bureau of Economic Geology. She has expertise in the areas of seismic site response analysis, engineering seismology, seismic slope stability, field reconnaissance after earthquakes, and remote sensing of geotechnical phenomena. Dr. Rathje is a founding member and current co-chair of the Geotechnical Extreme Events Reconnaissance (GEER) Association, and she was a member of the board of directors of the Earthquake Engineering Research Institute (EERI) from 2010-2013. She is the principal investigator for the DesignSafe-ci.org cyberinfrastructure for the NSF-funded Natural Hazards Engineering Research Infrastructure (NHERI) and co-PI for the Center for Integrated Seismicity Research (CISR) at the Bureau of Economic Geology. She has been honored with various research awards, including the Huber Research Prize from the American Society of Civil Engineers (ASCE) in 2010, the Hogentogler Award for outstanding paper from ASTM Committee D18 in 2010, the Shamsher Prakash Research Award in 2007, and the Shah Innovation Prize from EERI in 2006. She was named a fellow of the American Society of Civil Engineers in 2016.

Appendix C. List of Attendees

Note: * indicates participants who registered but could not attend".

First Name	Last Name	Affiliation	Email
Andrew	Adamson	Gemini Observatory	aadamson@gemini.edu
Tim	Ahern	Incorporated Research Institutions for Seismology (IRIS)	tim@iris.washington.edu
**James	Allen	National Science Foundation	jallan@nsf.gov
Aaron	Andersen	National Center for Atmospheric Research	aaron@ucar.edu
Stuart	Anderson	Caltech	stuart.anderson@ligo.org
Demian	Bailey	Oregon State University	baileyd@oregonstate.edu
Anjuli	Bamzai	National Science Foundation	*abamzai*@*nsf*.gov
Chaitan	Baru	National Science Foundation	baru@sdsc.edu
Stace	Beaulieu	Woods Hole Oceanographic Institution	stace@whoi.edu
**Steven	Berukoff	AURA/DKIST	sberukoff@nso.edu
Fran	Boler	UNAVCO	fboler@unavco.org
Adam	Bolton	National Optical Astronomy Observatory	bolton@noao.edu
Devin	Bougie	CHESS - Cornell High Energy Synchrotron Source	devin.bougie@cornell.edu
Suzanne	Carbotte	Rolling Deck to Repository (R2R) Program	carbotte@ldeo.columbia.edu
**Kalyana	Chadalavada	Intel	Kalyana.chadalavada@intel.com
**Tim	Cockerill	University of TX	cockerill@tacc.utexas.edu
Mark	Coles	National Science Foundation	
Peter	Couvares	LIGO Laboratory - Caltech	peter.couvares@ligo.org
Eric	Cross	National Science Foundation - DKIST	ecross@nso.edu
Peter	Darch	University of Illinois at Urbana-Champaign	ptdarch@illinois.edu
Ewa	Deelman	University of Southern California	deelman@isi.edu
Thomas	DeFanti	UC San Diego	tdefanti@ucsd.edu
Dan	Fay	Microsoft Research	danf@Microsoft.com
Rafael	Ferreira da Silva	University of Southern California	rafsilva@isi.edu
Amy	Friedlander	National Science Foundation	afriedla@*nsf*.gov
**Nathan	Galli	University of Utah	nathang@sci.utah.edu
Rob	Gardner	University of Chicago	rwg@uchicago.edu
Philip	Gates	International Ocean Discovery Program	gates@iodp.tamu.edu
Forough	Ghahramani	Rutgers Informatics Institute	forough.ghahramani@rutgers.edu
Brian	Glendenning	National Radio Astronomy Observatory	bglenden@nrao.edu

Iain	Goodenow	Large Synoptic Survey Telescope	igoodenow@lsst.org
Bret	Goodrich	National Solar Observatory / Daniel K. Inouye Solar Telescope (NSO/DKIST)	goodrich@nso.edu
Raphael	Greenbaum	NHERI EF Wall of Wind	rgreenba@fiu.edu
Tom	Gulbransen	Battelle - NEON	gulbran@battelle.org
David M.	Halstead	National Radio Astronomy Observatory	dhalstead@nrao.edu
Rob	Hengst	National Science Foundation BFA/LFO	rhengst@nsf.gov
Pamela	Hill	National Center for Atmospheric Research	pjg@ucar.edu
**Bob	Houtman	National Science Foundation	bhoutman@nsf.gov
Julio	Ibarra	Florida International University	julio@fiu.edu
Margaret Gelman	Johnson	NCSA - DES/Large Synoptic Survey Telescope	mgelman2@illinois.edu
Greg	Jones	University of Utah Scientific Computing and Imaging Institute	greg@sci.utah.edu
Jeffrey	Kantor	AURA/Large Synoptic Survey Telescope	jkantor@lsst.org
Kate	Keahey	Mathematics & CS Division, Argonne National Laboratory/Computation Institute, University of Chicago	keahey@anl.gov
Ken	Klingenstein	Internet2	kjk@internet2.edu
Albert	Lazzarini	LIGO Laboratory, Caltech	lazz@ligo.caltech.edu
**Kerstin	Lehnert	Columbia University	lehnert@ldeo.columbia.edu
Elise	Lipkowitz	National Science Foundation	elipkowi@nsf.gov
Miron	Livny	University of Wisconsin-Madison	miron@cs.wisc.edu
Vyacheslav (Slava)	Lukin	National Science Foundation	vlukin@*nsf*.gov
Pedro	Marronetti	National Science Foundation	pmarrone@nsf.gov
James A.	Marsteller	Pittsburgh Supercomputing Center/CMU - CTSC/CCoE	jam@psc.edu
Aaron	Matthews	Northwestern University	aaron.matthews1@northwestern.edu
Sean	McManus	National Optical Astronomy Observatory	mcmanus@noao.edu
Gonzalo	Merino	University of Wisconsin Madison	gonzalo.merino@icecube.wisc.edu
Jon C.	Meyer	UC San Diego/Scripps Institution of Oceanography	jmeyer@ucsd.edu
Bogdan	Mihaila	National Science Foundation	bmihaila@*nsf*.gov
William	Miller	National Science Foundation	wlmiller@*nsf*.gov
Subhashree (Shree)	Mishra	National Science Foundation	sumishra@nsf.gov
Inder	Monga	Energy Sciences Network	imonga@es.net
Chris	Morrison	Gemini Observatory	cmorrison@gemini.edu
**Azad	Naeemi	Georgia Institute of Technology	azad@gatech.edu
Anita	Nikolich	National Science Foundation	anikolic@nsf.gov
Manish	Parashar	Rutgers Discovery Informatics Institute	parashar@rutgers.edu
**Joseph	Paris	Northwestern University	j-paris@northwestern.edu

Valerio	Pascucci	Center for Extreme Data Management, Analysis and Visualization	pascucci@sci.utah.edu
Stephanie	Petillo	Woods Hole Oceanographic Institution	spetillo@whoi.edu
Benoît	Pirenne	Ocean Networks Canada	bpirenne@oceannetworks.ca
Beth	Plale	National Science Foundation	bplale@nsf.gov
Sushil	Prasad	National Science Foundation	*sprasad*@*nsf*.gov
Phil	Puxley	National Science Foundation	ppuxley@nsf.gov
Irene	Qualters	National Science Foundation	iqualter@nsf.gov
Mohan	Ramamurthy	UCAR/Unidata Data Services and Tools for Geoscience	mohan@ucar.edu
**Arcot	Rajasekar	University of No. Carolina at Chapel Hill	rajasekar@unc.edu
Ellen	Rathje	University of Texas--DesignSafe CI	e.rathje@mail.utexas.edu
Karyn	Roberts		
Tom	Rockwell	NSCL/MSU National Superconducting Cyclotron Laboratory Michigan State University	rockwell@nscl.msu.edu
Ivan	Rodero	Rutgers Discovery Informatics Institute	irodero@rutgers.edu
Christopher	Romsos	Oregon State University, College of Earth, Ocean, and Atmospheric Sciences, Regional Class Research Vessel Program	cromsos@coas.oregonstate.edu
Jim	Rosser	International Ocean Discovery Program at Texas A&M University	jrosser@tamu.edu
Eric	Saltzman	National Science Foundation	
**Jim	Shank	NSF	
Adam	Shepherd	Biological & Chemical Oceanography Data Management Office	ashepherd@whoi.edu
Dan	Stanzione	Texas Advanced Computing Center, The University of Texas at Austin	arleen@tacc.utexas.edu
Marc	Stieglitz	National Science Foundation	mstiegli@nsf.gov
Victoria	Stodden	University of Illinois at Urbana-Champaign	vcs@stodden.net
Andreas	Stolz	National Superconducting Cyclotron Laboratory / Michigan State University	stolz@nscl.msu.edu
Werner	Sun	CHESS, Cornell University	wms8@cornell.edu
**Alexander	Szalay	Johns Hopkins University	szalay@jhu.edu
Guebre	Tessema	National Science Foundation	gtessema@*nsf*.gov
Kevin	Thompson	US National Science Foundation	kthompso@nsf.gov
John	Towns	National Center for Supercomputing Applications/Extreme Science and Engineering Discovery Environment	admoore@illinois.edu
**Steve	Tuecke	Globus U of Chicago	tuecke@uchicago.edu
Rick	Wagner	Globus	rick@globus.org
**Amy	Walton	NSF	awalton@nsf.gov
Yuanxi	Wang	2-Dimensional Crystal Consortium, Pennsylvania State University	yow5110@psu.edu

Frank	Wuerthwein	UCSD/San Diego Supercomputer Center	fkw@ucsd.edu
Michael	Zentner	Purdue University / HUBzero	mzentner@purdue.edu

Appendix D. Pre-workshop Survey Responses

Pre-workshop Survey Questionnaire

The pre-workshop survey questionnaire consisted of the following questions:

- What significant components of the CI were developed in-house? Are these components available to others to reuse?
- What external CI capabilities and services and/or externally developed tools (if any) does the facility use and who provides them? How were these tools identified and what criteria were used to select the tools?
- List up to 3 of your most used and most challenging CI components with a 1-sentence explanation for each. What aspects of the facility CI and its operation would you like to share as best practices?
- What aspects of the facility CI and its operation do you see as challenges or gaps? Are there “CI lessons learned” that you would like to share or see discussed at the workshop?
- What do see you as key risks in facility CI (e.g., dependency on external resources (compute, data, expertise) and/or services)? Are there mitigation steps that you would like to share or see discussed at the workshop?
- What CI-related workforce development activities does your facility engage in?
- What do you see as your key new CI requirements and challenges in the next 5-10 years?
- Do you have any other suggestions for the workshop?

Affiliation	Name	E-mail
UCAR/Unidata	Mohan Ramamurthy	mohan@ucar.edu
What percentage of the facility CI was developed in-house versus by reusing existing solutions?		
Data systems and services, software/middleware and tools; Almost all data and software from Unidata are made available freely and openly and use open source licensing, so they can be reused.		
What external CI capabilities and services and/or externally developed tools (if any) does the facility use and who provides them? How were these tools identified and what criteria was used to select the tools?		
In addition to Unidata-developed software, we also provide externally developed software to our users. Such tools are identified based on the needs of the academic users and deliberated by our governing committees.		
List up to 3 of your most and least favorite CI components with a 1 sentence explanation for each. What aspects about the facility CI and its operation would you like to share as best practices?		
NetCDF is Unidata's most widely used software. The challenge is to provide support to a very large and diverse user base in almost every country in the world and all geoscience domains and sectors. The Local Data Manager and THREDDS Data Server applications also have a diverse user community in both operational and research settings. Providing support to an ever expanding community remains an ongoing challenge. Another challenge stems from the rapid growth in the volume of data, so a push approach will not not be sustainable. The increasing volume and diversity of data sources, coupled with the growing user base, also creates challenges in scaling and interoperability.		
What aspects of the facility CI and its operation do you see as challenges/gaps? Are there any pitfalls/mistakes you would like to share? What aspects would you be interested in outsourcing?		
As stated earlier, maintaining high quality of support to a growing and expanding user base in an era of shrinking or level budgets remains a challenge. There are also sociological and cultural challenges with changing technologies and adoption and use of new tools and services. Migration to cloud platforms poses challenges in developing business and cost recovery models.		
Key Risks		
The lack of NSF-funded operational cloud facilities for hosting data and delivering services remains a key gap. Also, most CI facilities are operating independently without much		

collaboration and partnership. In addition to sharing knowledge and expertise, a discussion on how the facilities can share other resources and infrastructure would be valuable.
What CI-related workforce development activities does your facilities engage in?
Unidata provides education and training, through workshops in Boulder and at different universities, on a regular basis to students and faculty on its products and services. In addition, Unidata hosts several interns and mentors them every summer.
What do you see as your key new CI requirements and challenges in the next 5-10 years
Exploding data volumes and scaling of CI to meet the growing needs remains a challenge. Cybersecurity is another challenging area. Entraining and retaining professionals into scientific CI areas is a challenge given that graduating students and professionals are paid much more by the IT and software industry that is thriving.
Do you have any other suggestions for the workshop?
Clearly stated goals for the workshop and more in-depth discussions on important issues (rather than many overview presentations) is likely to lead to meaningful outcomes.

Affiliation	Name	E-mail
NEON	Tom Gulbransen, Battelle	gulbransen@battelle.org
What percentage of the facility CI was developed in-house versus by reusing existing solutions?		
3 ingestion queues, 4 transformation pipelines, 2 websites. Tailored so unlikely to reuse.		
What external CI capabilities and services and/or externally developed tools (if any) does the facility use and who provides them? How were these tools identified and what criteria was used to select the tools?		
6 external host partners for community distribution and limited data product creation. AeroNet, MG-Rast, SRA, BOLD, PhenoCam, AmeriFlux		
List up to 3 of your most and least favorite CI components with a 1 sentence explanation for each. What aspects about the facility CI and its operation would you like to share as best practices?		
Sensor messaging and control challenging at sites infrequently visited. Ingestion queues which can accommodate dozens of data types and sources. APIs which greatly simply powerful data access and sharing options.		
What aspects of the facility CI and its operation do you see as challenges/gaps? Are there any pitfalls/mistakes you would like to share? What aspects would you be interested in outsourcing?		
The fusion of classical IT systems development now in ntegralky relies on code written by non-IT analysts. The value of the latter was underestimated initially, and will be over-emphasized going forward during community engagement.		
Key Risks		
Sensor unreliability is a risk addressed by engineering. User diversity will create demands beyond the dev team capacity. Initial Ops period will reveal if/where/when/how cyberinfrastructure may need to automate more checks and editsbility.		
What CI-related workforce development activities does your facilities engage in?		
Lots of cyberinfrastructure recruitment and resultant learning curve climbing during construction. Scientific cosers are being herded toward conventions to promote easier		

interoperability and expansion through external contributions which can be evaluated.
What do you see as your key new CI requirements and challenges in the next 5-10 years
User community traceability and expansion of user's demands.
Do you have any other suggestions for the workshop?
Share registrants info.

Affiliation	Name	E-mail
Ocean Observatory Initiative (OOI)	Ivan Rodero, Rutgers University	irodero@rutgers.edu
What percentage of the facility CI was developed in-house versus by reusing existing solutions?		
The infrastructure of the CI has been developed in-house following industry best practices. It includes the data lifecycle management system, and the network and system architecture distributed across two geographically distributed data centers. The customized software stack, including core data management system and user interface has been also developed. The CI architecture and best practices are available to other to reuse.		
What external CI capabilities and services and/or externally developed tools (if any) does the facility use and who provides them? How were these tools identified and what criteria was used to select the tools?		
The OOI CI uses a number of external services and tools, including an Apache server for raw data delivery, a THREEDS server for asynchronous data product delivery, Alfresco for document configuration management and shipboard data delivery, and a number of tools such Redmine and Confluence for documentation and configuration management, gerrit and Jenkins for continuous integration, and phpBB for forums. These tools were selected based on requirements and prioritizing open source solutions, when needed.		
List up to 3 of your most and least favorite CI components with a 1 sentence explanation for each. What aspects about the facility CI and its operation would you like to share as best practices?		
1) On-demand data product delivery: OOI provides users with a graphical user interface (i.e., OOINet data portal) for plotting and downloading on-demand data products. The portal also provides access to live video and other data products. 2) Raw data archive: data is available for download in “raw” indicates data as they are received directly from the instrument, in instrument-specific format. 3) Machine-to-machine API: a REFTful user interface is available to access OOI CI programmatically using authentication mechanisms. We’d like to share the architecture of the enterprise-level information lifecycle management system, including networking and monitoring components which use industry best practices.		
What aspects of the facility CI and its operation do you see as challenges/gaps? Are there		

any pitfalls/mistakes you would like to share? What aspects would you be interested in outsourcing?
Two of the most important challenges of the OOI CI are 1) evolving requirements (e.g., data rates, services), 2) and integration of new components (e.g., new instruments). There are lessons learnt related to the implementation of industry best practices for the deployment and operation of a production-level CI.
Key Risks
One of the highest risks for the OOI CI is related to the uncertainties for keeping the funding level for operating and maintaining the core infrastructure, the software stack and fundamental services. For example, the lack of expanding the storage infrastructure in the future is a risk. A mitigation step was including expandable tape-base storage infrastructure in the information lifecycle management system.
What CI-related workforce development activities does your facilities engage in?
CI-related workforce development is at different levels. On the one hand, technical personnel are engaged with continuous training on the technologies involved in CI (e.g. Palo Alto training, Dell Compellent, Apache Cassandra, etc.). On the other hand, OOI engaged with NSF-funded CTSC for the development of a comprehensive cyber-security plan.
What do you see as your key new CI requirements and challenges in the next 5-10 years
New CI requirements/challenges in the next 5-10 are related to the expansion of the CI network with new instruments, increasing data rates and evolving data delivery mechanisms.
Do you have any other suggestions for the workshop?
Not at this time

Affiliation	Name	E-mail
National Nanotechnology Coordinated Infrastructure (NNCI)	Azad Naeemi, Georgia Institute of Technology	azad@gatech.edu
What percentage of the facility CI was developed in-house versus by reusing existing solutions?		
Institute developed components include a self-service firewall management, and a shared access model where institute purchased equipment is provided to faculty who in return provide shared access to their purchased hardware.		
What external CI capabilities and services and/or externally developed tools (if any) does the facility use and who provides them? How were these tools identified and what criteria was used to select the tools?		
We are actively implementing the Open Science Grid, Globus, science DMZ, and perfSONAR file and networking components. In addition, we are implementing Ohio Supercomputing Center's PBS Tools, Open XDMoD from the University at Buffalo.		
List up to 3 of your most and least favorite CI components with a 1 sentence explanation for each. What aspects about the facility CI and its operation would you like to share as best practices?		
1) Rapidly growing data sources. Our storage systems have grown exponentially since 2009 to 8 petabytes. 2) Utilization patterns that are many small jobs, i.e. high throughput computing (HTC) vs the few very large monolithic jobs (HPC). We aim to funnel these types of workloads to OSG, and implement hardware dedicated to running OSG computation.		
What aspects of the facility CI and its operation do you see as challenges/gaps? Are there any pitfalls/mistakes you would like to share? What aspects would you be interested in outsourcing?		
Key Risks		
Not at this time		

What CI-related workforce development activities does your facilities engage in?
We hire undergraduate students, contribute to Linux Cluster Institute workshops and are in the process of deploying an instructional cluster.
What do you see as your key new CI requirements and challenges in the next 5-10 years
As a major technological research institution, the Georgia Institute of Technology, which includes academic units and the Georgia Tech Research Institute (GTRI), has direct experience with many of the current and emerging research challenges facing today's
Do you have any other suggestions for the workshop?
Not at this time

Affiliation	Name	E-mail
NHERI	Tim Cockerill, University of Texas - Texas Advanced Computing Center	cockerill@tacc.utexas.edu
What percentage of the facility CI was developed in-house versus by reusing existing solutions?		
Nearly all of the CI components are developed in-house by TACC and are made available as open source in github.		
What external CI capabilities and services and/or externally developed tools (if any) does the facility use and who provides them? How were these tools identified and what criteria was used to select the tools?		
We use the Django web framework based on our previous experiences with this and other frameworks. We also have a local implementation of the Fedora Digital Object Repository Management System for our archiving our published data.		
List up to 3 of your most and least favorite CI components with a 1 sentence explanation for each. What aspects about the facility CI and its operation would you like to share as best practices?		
The Data Depot is our most used CI component. Our users have already uploaded more than 16TB of data in addition to the 40TB we transitioned in from the predecessor project NEES. We allow all file types and we encourage our users to upload any and all data they need to do their research - we feel that not restricting the users is key to their adoption of our CI. We worked with Mathworks to acquire a MATLAB license that enables all academic users to access MATLAB via our CI. The engineering community are heavy MATLAB users, and this has also helped with adoption. We implemented Jupyter Notebooks and are providing training on how to use them along with basic Python scripting skills. We are seeing pretty strong uptake of Jupyter. It runs pretty fast in the cloud, and users are finding it to be as capable as MATLAB.		
What aspects of the facility CI and its operation do you see as challenges/gaps? Are there any pitfalls/mistakes you would like to share? What aspects would you be interested in outsourcing?		
Challenge: operation of a tightly-coupled operation across hemispheres		

It is preliminary to speak of lessons lesson learned, as LSST is in construction. However, accurate and detailed model to effectively communicate, coordinate and maintain the ability to trace CI features to the requirements and business need. Is an area of focus which LSST feels will help meet this challenge.
Key Risks
For this project, since the CI is all at TACC, there is not much risk.
What CI-related workforce development activities does your facilities engage in?
We provide roughly monthly training webinars which are recorded and then made available persistently on YouTube. We also have summer programs for high school students - this year they built an instrumented model, experimented with that model on a shake table, and then analyzed their results using our CI.
What do you see as your key new CI requirements and challenges in the next 5-10 years
Performance is the priority, since web data transfer and remote use of interactive tools like MATLAB are slower than on a local laptop. Also expanded simulation and data analysis/visualization capabilities on the web portal so that we capture all researchers in this community.
Do you have any other suggestions for the workshop?
Not at this time

Affiliation	Name	E-mail
LSST	Don Petraivck, NCSA - UIUC Jeff Kantor, William O'Mullane	Petravick@illinois.edu
What percentage of the facility CI was developed in-house versus by reusing existing solutions?		
R: LSST is in construction, but the following are underway, LSST has funded the development of a significant, high bandwidth network between Chile and the United States. LSST is developing QSERV, a spatially shared database which is anticipated to require 40 PB of disk provisioning, over 250 node by 2025.		
What external CI capabilities and services and/or externally developed tools (if any) does the facility use and who provides them? How were these tools identified and what criteria was used to select the tools?		
- LSST Uses HT-CONDOR for the basis of its production system. HT-Condor is a standard in thoughput computing, is used in LHC and the Dark Energy survey. HTCondor supports the various batch use cases identified in LSST. LSST has had a collaborative engagement with HTCondor for many years. LSST has used XSEDE and Blue Waters during its pre-construction phase for demonstrations of feasibility of its production system, and has used simulation data generated on the Open Science Grid. – These were the obvious choices dues to agency support and availability. LSST has built upon authentication and authorization system work that is also in use in LIGO. The reason is that the system supports a variety of authentication and authorization protocol, and interoperated with Incommon. National education and research identity federations are seen as useful source of identity information for LSST, where the class of all US and all Chilean professional astronomers have data rights. LSST's Master Information Security Plan was developed in Consultation with the CTSC. CTSC was selected due it is knowledge of contemporary security standards, as applied to NSF projects. LSST's science user interface is based on the Firefly Tool Kit developed at IPAC at Caltech. This is a commonly used advanced toolkit used within Optical Astronomy. Rucio, a component developed at CERN for the LHC is being evaluated for internal file synchronization, as is Pegasus for the production workflows. Both of these components were selected due to their use with similar use cases in other experiments.		

Jupyter is a foundational component to support internal quality assessment and to support exploitation of the data at the UN and Chilean LSST Data Access Centers. Jupyter is a well-supported method of exposing aspects of a facility in a structured way to a large group of users.

BRO is use for intrusion detection at the LSST Chilean sites, and at NCSA. BRO is selected for us utility in being an intrusion detection system where large volumes of data re transferred between sites, and sue to the body of expertise with the system at NCSA

List up to 3 of your most and least favorite CI components with a 1 sentence explanation for each. What aspects about the facility CI and its operation would you like to share as best practices?

- LSST Uses HT-CONDOR for the basis of its production system. HT-Condor is a standard in throughout computing, is used in LHC and the Dark Energy survey. HTCondor supports the various batch use cases identified in LSST. LSST has had a collaborative engagement with HTCondor for many years.

LSST has used XSEDE and Blue Waters during its pre-construction phase for demonstrations of feasibility of its production system, and has used simulation data generated on the Open Science Grid. – These were the obvious choices dues to agency support and availability.

LSST has built upon authentication and authorization system work that is also in use in LIGO. The reason is that the system supports a variety of authentication and authorization protocol, and interoperated with Incommon. National education and research identity federations are seen as useful source of identity information for LSST, where the class of all US and all Chilean professional astronomers have data rights.
LSST's Master Information Security Plan was developed in Consultation with the CTSC. CTSC was selected due it is knowledge of contemporary security standards, as applied to NSF projects.
LSST's science user interface is based on the Firefly Tool Kit developed at IPAC at Caltech. This is a commonly used advanced toolkit used within Optical Astronomy.
Rucio, a component developed at CERN for the LHC is being evaluated for internal file synchronization, as is Pegasus for the production workflows. Both of these components were selected due to their use with similar use cases in other experiments.
Jupyter is a foundational component to support internal quality assessment and to support exploitation of the data at the UN and Chilean LSST Data Access Centers. Jupyter is a well-supported method of exposing aspects of a facility in a structured way to a large group of users.
BRO is use for intrusion detection at the LSST Chilean sites, and at NCSA. BRO is selected for us utility in being an intrusion detection system where large volumes of data re transferred between sites, and sue to the body of expertise with the system at NCSA

1) Upgrading the north south network from LaSerena, Chile to NCSA in the context of a MREFC project.
2) Dealing with the evolution of processors, in particular the reduction of the amount of memory per core, and the need to increase the level of threading in LSST Codes.
3) Selecting the technologies needed to support end users in the data access center.

What aspects of the facility CI and its operation do you see as challenges/gaps? Are there any pitfalls/mistakes you would like to share? What aspects would you be interested in outsourcing?

Challenge: operation of a tightly-coupled operation across hemispheres
It is preliminary to speak of lessons lesson learned, as LSST is in construction. However, accurate and detailed model to effectively communicate, coordinate and maintain the ability to trace CI features to the requirements and business need. Is an area of focus which LSST feels will help meet this challenge.

Key Risks

Changes in computing platforms over the remaining period of construction and operations through 2034 are a concern. LSST has data processing access and archive facilities in three continents. For each continent the pace of sustainable change will vary. For example, we expect cloud computing to lag in South America. The response to these challenges includes providing software isolation layers, for example Kubernetes, which can be deployed in locally provisioned or in commercial systems.

We currently use could services for software build and test. The EPO component of LSST has a very large cloud deployment component. Our baseline thinking allows for use of cloud services for disaster recovery, for opportunistic bulk computing, and for elastic expansion of the US Data Access centers. Our baseline may evolve as construction proceeds.

What CI-related workforce development activities does your facilities engage in?

Project staff attend workshops and conferences. At NCSA significant work in CI is performed by NCSA staff. NCSA has a program of work to develop the HPC workforce, including responding to NSF calls for proposals for training Cyber Infrastructure Professionals. Additionally, NCSA has a program of research and supporting its infrastructure, including operational security group, support for the Linux Cluster Institute (LCI), which trains Infrastructure professionals.

What do you see as your key new CI requirements and challenges in the next 5-10 years

Keeping the CI efforts in Chile and the in the US coordinated and with a like technology base. Changes in CI technologies and how CI is absorbed by the project. LSST has obligations to provide computing facilities in Chile, where for example cloud functionality is not equivalent to the functionality available in the US.
Do you have any other suggestions for the workshop?
Not at this time

Affiliation	Name	E-mail
National Optical Astronomy Observatory (NOAO)	Sean McManus	mcmanus@noao.edu

What percentage of the facility CI was developed in-house versus by reusing existing solutions?
data reduction pipeline (DEC Community Pipeline); TADA (Telescope Automatic Data Archiver); yes these tools are mostly open-source
What external CI capabilities and services and/or externally developed tools (if any) does the facility use and who provides them? How were these tools identified and what criteria was used to select the tools?
Scientific Linux, IBM General Parallel File System, Puppet, Foreman, Libvirt, Django. The criteria used to select tools varies. For some open-source tools, there is minimal investment needed to try something, and therefore doesn't require a formal selection process. For paid software contracts, there is obviously more vetting by internal IT staff, management, and procurement. As part of normal vetting we try to look at what is working / not working for other peer organizations inside and outside of AURA.
List up to 3 of your most and least favorite CI components with a 1 sentence explanation for each. What aspects about the facility CI and its operation would you like to share as best practices?
1) Mass storage: We require inexpensive storage on the multi-Petabyte scale to store astronomy data products; 2) Bandwidth: Reliable, fast bandwidth across continents is needed to move data from telescope to archive; 3) Software: The software stack must meet operational requirements but also be sustainable inside flat or shrinking budget envelope.
What aspects of the facility CI and its operation do you see as challenges/gaps? Are there any pitfalls/mistakes you would like to share? What aspects would you be interested in outsourcing?
For small departments, it is difficult to achieve a balance of experience versus motivation and familiarity with cutting edge tools. Low staff turnover can result in staff being settled on one particular technology, and lagging behind recent developments in IT. On the other hand, it's not cost-effective to react to the latest/greatest thing that comes out every year.

A balance of new versus proven tools must be made.
Key Risks
workforce reduction due to budgets, even a small one, could have significant impact.
What CI-related workforce development activities does your facilities engage in?
We budget for continuing education, but whether or not staff participate is voluntary
What do you see as your key new CI requirements and challenges in the next 5-10 years
transition from NOAO/LSST/Gemini to NCOA
Do you have any other suggestions for the workshop?
n/a

Affiliation	Name	E-mail
LIGO	Stuart Anderson, Caltech	stuart.anderson@ligo.org

What percentage of the facility CI was developed in-house versus by reusing existing solutions?

All of the following in-house CI components are available for reuse:
* LIGO Data Replicator (bulk data transfers)
* Metadata databases and tools designed for GW observations
* low-latency data distribution on large clusters
* Data Monitoring Tools
* low-latency transient event alert system
* Network Data Server
* Web and Matlab based Data Viewer tools
* GW Detector status monitoring service
* GW detection and parameter estimation pipelines
* Library of gravitational wave algorithms
* LIGO Open Science Center notebooks
* Job accounting system

What external CI capabilities and services and/or externally developed tools (if any) does the facility use and who provides them? How were these tools identified and what criteria was used to select the tools?

* HTCondor/Pegasus/BOINC
* OSG
* Docker/Singularity/Shifter
* CVMFS/StashCache/Xrootd/GridFTP
* Shibboleth/Grouper/CILogon/Kerberos/LDAP/GSI
* Oracle HSM/ZFS/HDFS
* GitHub/GitLab/Travis/Jenkins
* JupyterHub

These tools where predominantly identified by first recognizing a need and then charging a small group to research (sometimes a self-forming group) to research what is currently available. In some cases that group takes a solution to full scale prototype (build it and they will come), and in others the alternatives are presented to a LIGO computing committee to evaluate the pros and cons first.and Matlab based Data Viewer tools*GW Detector status monitoring service*GW detection and parameter estimation pipeline*Library of gravitational wave algorithms*LIGO Open Science Center notebooks*Job accounting system

List up to 3 of your most and least favorite CI components with a 1 sentence explanation for each. What aspects about the facility CI and its operation would you like to share as best practices?
* Identity and Access Management was a challenge during the early phases of LIGO, leading to significant loss in productivity due to unnecessary barriers to efficient access to needed information and systems. Integrating Shibboleth, Grouper, InCommon, and CILogon into LIGO's CI has been a game changer. Investing in I&AM early on in a project is highly recommended. * In the early years of LIGO attempts to use OSG to run LIGO data analysis tasks failed. In the last few years this has become a major success, in part due to more mature tools for managing data intensive workflows (e.g., Pegasus, CVMFS, and containerization), and in part due to more mature gravitational wave data analysis pipelines. * LIGO initially invested in a home grown job execution environment that attempted to minimize the amount of code needed to be developed by scientists performing searches for gravitational waves.. However, that proved in practice to be insufficiently flexible and the pendulum swung over to allowing scientists to develop arbitrary a.out executables managed by HTCondor. In hind site, the optimum would have been somewhere in-between.
What aspects of the facility CI and its operation do you see as challenges/gaps? Are there any pitfalls/mistakes you would like to share? What aspects would you be interested in outsourcing?
* Integrating CI with international collaborators remains a significant challenge.. OSG has recently provided a major breakthrough for providing a uniform interface to plan and execute LIGO workflows on international computing resources. However, international federated I&AM remains a significant challenge for LIGO. * Finding the right set of CI to support both tightly controlled production data analysis and allowing creative new ideas be developed is a challenge.
Key Risks
* Funding for CI experts that support scientific personnel to use existing CI * Sustainability of CI and being able to effectively identify new CI that will be available in the long-term before investing limited internal resources.
What CI-related workforce development activities does your facilities engage in?
* Sending students to summer schools and similar training opportunities. * Sending professional staff to conferences and workshops. * Inviting external experts to provide training at internal scientific meetings.

What do you see as your key new CI requirements and challenges in the next 5-10 years
* Inter-federation agreements that comply with international privacy laws while still releasing enough information to be useful for international scientific collaborations. * Training the teachers. As most of the workforce comes from academic research groups how do we train academic faculty to be able to train their new students to use modern CI. * long-term stability of software packaging and distribution that will allow reproducibility of scientific results on an interesting time scale.
Do you have any other suggestions for the workshop?
Not at this time

Affiliation	Name	E-mail
LIGO	Albert Lazzarini, Caltech	lazz@ligo.caltech.edu
What percentage of the facility CI was developed in-house versus by reusing existing solutions?		
Please see white paper submitted by Stuart Anderson for all attendees from LIGO		
What external CI capabilities and services and/or externally developed tools (if any) does the facility use and who provides them? How were these tools identified and what criteria was used to select the tools?		
Please see white paper submitted by Stuart Anderson for all attendees from LIGO		
List up to 3 of your most and least favorite CI components with a 1 sentence explanation for each. What aspects about the facility CI and its operation would you like to share as best practices?		
Please see white paper submitted by Stuart Anderson for all attendees from LIGO		
What aspects of the facility CI and its operation do you see as challenges/gaps? Are there any pitfalls/mistakes you would like to share? What aspects would you be interested in outsourcing?		
Please see white paper submitted by Stuart Anderson for all attendees from LIGO		
Key Risks		
Please see white paper submitted by Stuart Anderson for all attendees from LIGO		
What CI-related workforce development activities does your facilities engage in?		
Please see white paper submitted by Stuart Anderson for all attendees from LIGO		
What do you see as your key new CI requirements and challenges in the next 5-10 years		
Please see white paper submitted by Stuart Anderson for all attendees from LIGO		
Do you have any other suggestions for the workshop?		

What is the appropriate scale and relationship among large NSF computing facilities, computing facilities that are part of e.g., physics large facilities and MRI resources provided to individual collaboration institutions? Does NSF have a policy on these?

Affiliation	Name	E-mail
ARF	Jon C. Meyer, UC San Diego	jmeyer@ucsd.edU
What percentage of the facility CI was developed in-house versus by reusing existing solutions?		
we are in the process of developing data delivery via modern message queue and welcome the opportunity to collaborate and have others reuse.		
What external CI capabilities and services and/or externally developed tools (if any) does the facility use and who provides them? How were these tools identified and what criteria was used to select the tools?		
Some vendors' tools are used due the demand for certain types of data to be regularly produced during a seagoing mission		
List up to 3 of your most and least favorite CI components with a 1 sentence explanation for each. What aspects about the facility CI and its operation would you like to share as best practices?		
Uninterrupted Internet connectivity. Research vessels at sea need consistent, reliable communication paths to be able to produce scientifically interesting data in near to real time.		
What aspects of the facility CI and its operation do you see as challenges/gaps? Are there any pitfalls/mistakes you would like to share? What aspects would you be interested in outsourcing?		
Key Risks		
What CI-related workforce development activities does your facilities engage in?		
Some specialized and general computing-related training.		
What do you see as your key new CI requirements and challenges in the next 5-10 years		

High-speed, realtime delivery of data from the ocean. Ability to interact with field researchers seamlessly from
Do you have any other suggestions for the workshop?
Not at this time

Affiliation	Name	E-mail
Gemini	Chris Morrison, Gemini Observatory	cmorrison@gemini.edu

What percentage of the facility CI was developed in-house versus by reusing existing solutions?
none (note that we do not include software in our definition of CI)
What external CI capabilities and services and/or externally developed tools (if any) does the facility use and who provides them? How were these tools identified and what criteria was used to select the tools?
Google apps for business; Amazon web services; zoom conferencing services. Identified in all cases by industry surveys & best practices; selection via requirements analysis, in some cases usability analyses, and value for money.
List up to 3 of your most and least favorite CI components with a 1 sentence explanation for each. What aspects about the facility CI and its operation would you like to share as best practices?
Challenges: 1. Netapp storage. Large impact if this redundant system fails. 2. Backup storage infrastructure. Expensive, complex and requires significant expertise. 3. Remote access connectivity. Brings user management and security concerns. Best practices: 1. Gemini infrastructure has significant redundancy, as a result of lessons learned in previous failures. 2. Use of cloud service (AWS) for large-scale data archiving and access. 3. CI replacement policy on equipment at end of warranty.
What aspects of the facility CI and its operation do you see as challenges/gaps? Are there any pitfalls/mistakes you would like to share? What aspects would you be interested in outsourcing?
Challenges & gaps: see above. Lessons to share: Redundancy (storage, networking, VM clusters, connectivity). Lessons to learn in the meeting: offsite storage methods & data retention.

Key Risks
Dependencies: Access to Google (for business applications); AWS (for archive storage) - low likelihood, high impact risks. Mitigation: Redundant network links in Hawaii and Chile. Backup plan for an extended outage of AWS would be to bring the archive in house temporarily until service restored.
What CI-related workforce development activities does your facilities engage in?
Enterprise specialist training courses and certifications.
What do you see as your key new CI requirements and challenges in the next 5-10 years
Challenge: Integration of Gemini CI into a larger Center, and aligning services with other Programs in that Center. We do not see significant changes in the technical challenge for Gemini CI, as the telescopes will not fundamentally change the way they operate at night.
Do you have any other suggestions for the workshop?
1. Future role of NSF in coordinating or providing CI through grant funding. 2. Large-scale science data storage and access via cloud services - best practices.

Affiliation	Name	E-mail
DKIST, NSO	Steve Berukoff and Eric Cross, NSO	sberukoff@nso.edu ecross@nso.edu

What percentage of the facility CI was developed in-house versus by reusing existing solutions?

For the DKIST telescope
Built In-House
- Instrument Control Systems
- Facility Control Systems
- Telescope
- Enclosure
- Environmental
- Adaptive Optics, Wavefront Control
- Coude
- Safety Systems
- Are these useful to other CI organizations?

Unclear if they would be useful elsewhere.

What external CI capabilities and services and/or externally developed tools (if any) does the facility use and who provides them? How were these tools identified and what criteria was used to select the tools?

- Open Source software; given budgetary constraints DKIST CI is leveraging Open Source where applicable. The deployment of Open Source is centered within the Infrastructure layers.
- Globus GridFTP will be ustilized to move data from the telescope on Maui to the Boulder Data Center.
- CEPH object storage for long-term data storage

List up to 3 of your most and least favorite CI components with a 1 sentence explanation for each. What aspects about the facility CI and its operation would you like to share as best practices?

- Complexity of DKIST Instruments has driven a flexible but customizable approach to instrument controls.
- Data network management has provided a challenge to DKIST. We have network Interconnects between the DKIST Facility on Maui, the University of Hawaii, the University of Colorado, and also leveraging Internet2.
- Complexity of DKIST Instruments has driven a flexible but customizable approach to

instrument controls. • Data network management has provided a challenge to DKIST. We have network Interconnects between the DKIST Facility on Maui, the University of Hawaii, the University of Colorado, and also leveraging Internet2. • The combination of Petascale data volume under a very constrained budget challenges the ability of the CI to support its community. Best Practices • Because of the distributed nature of the program with multiple product owners following Systems Engineering practices for developing effective requirements and interface controls.
What aspects of the facility CI and its operation do you see as challenges/gaps? Are there any pitfalls/mistakes you would like to share? What aspects would you be interested in outsourcing?
• Ensuring the end to end CI design from Facility Control, Data Acquisition and end-user distribution is built-in to the overall design and budget.
Key Risks
• Operational funding levels should allow appropriate maintenance to be completed with appropriate personnel. • Long-Term operational lifetimes mandate avoidance of monolithic architectures. Mitigation • Ability to build infrastructure building blocks by developing a roadmap for DIBBS awards.
What CI-related workforce development activities does your facilities engage in?
• Professional development conferences
What do you see as your key new CI requirements and challenges in the next 5-10 years
• Ensure we can deliver the scope that we need to support our community.
Do you have any other suggestions for the workshop?
Not at this time

Affiliation	Name	E-mail
ARF	Suzanne Carbotte, Columbia University	carbotte@ldeo.columbia.edu

What percentage of the facility CI was developed in-house versus by reusing existing solutions?
R2R has developed a network file system for storage of data and documents; a relational database for storage of associated metadata; a Web portal for search, browse, and download; scripted tools for data cataloging, archiving, processing, and assessment; and a suite of Web services for interoperability. Most are built on existing open-source software such as PostgreSQL, Apache HTTP/Tomcat, MapServer, etc. Selected tools for data processing have been released in the public domain via GitHub.
What external CI capabilities and services and/or externally developed tools (if any) does the facility use and who provides them? How were these tools identified and what criteria was used to select the tools?
R2R uses commercial provisioning in selected cases for Web service hosting (Linode.com), domain services (Site5.com), and deep storage (Amazon Glacier).
List up to 3 of your most and least favorite CI components with a 1 sentence explanation for each. What aspects about the facility CI and its operation would you like to share as best practices?
1. R2R's network file system is the heart of its daily operation, used for both internal processing workflows and serving content to the Web. The file system is built on a suite of FibreChannel storage arrays, switches, and Linux servers. 2. R2R's "NavManager" software package is used routinely to create a suite of quality-controlled shiptrack navigation products, which are reused by downstream QA processes and Web services. 3. R2R's "Linked Data" server disseminates the Cruise Catalog in a standards-compliant format, which is harvested by other geoscience data repositories as well as by global search indexes such as Google. What aspects about the facility CI and its operation would you like to share as best practices? It is not uncommon to revisit old(er) data packages, in order to extract additional information and/or refine quality assessment. Maintaining data packages on spinning disk for a 5 or more-year sliding window has proven advantageous, and can be sustained using (less expensive) HDDs rather than SSDs. Every digital resource published online (vessel, cruise, dataset, document, sample, person, award, etc) should have a globally unique persistent identifier. This enables interoperability

with other repositories, reliable citation, and linking to the scientific literature.
What aspects of the facility CI and its operation do you see as challenges/gaps? Are there any pitfalls/mistakes you would like to share? What aspects would you be interested in outsourcing?
The volume of environmental sensor data being produced by modern research vessels, is increasing faster than the disk storage capacity that can be deployed with affordable enterprise-grade local equipment. Commercial provisioning provides an affordable solution for deep storage, but not for local data processing or egress. Academic provisioning via systems like XSEDE is difficult because the resources are disjointed and constantly evolving, and carry the risk of abrupt termination when the grant period ends. Data transfer is also hampered by local campus network bandwidth. While progress has been made toward standardization, the US. academic fleet still produces data in a very heterogeneous manner. Each cruise is unique. Significant manpower is still required to stay abreast of changing directory structures and file formats, and to recover from operator errors.
Key Risks
Maintaining local server, storage, and network infrastructure remains an ongoing challenge, especially with the increased need to provide monitoring, metrics, and network security. Commercial provisioning shifts resources from a local to a remote location, but does not eliminate the need for a system administrator and does not reduce costs.
What CI-related workforce development activities does your facilities engage in?
R2R staff attend annual community meetings such as ESIP, RDA, and RVTEC, to stay abreast of emerging technologies. Junior staff work in tandem with senior staff, receiving on-the-job training.
What do you see as your key new CI requirements and challenges in the next 5-10 years
The ability to store and move large volumes of data as environmental sensors continue to evolve faster than storage/network resources; the lack of "smart" self-documenting sensors; and the lack of designated long-term archives for some data types remain significant challenges.
Do you have any other suggestions for the workshop?
Not at this time

Affiliation	Name	E-mail
National Center for Atmospheric Research (NCAR)	Aaron Andersen, UCAR	aaron@ucar.edu

What percentage of the facility CI was developed in-house versus by reusing existing solutions?

A number of components of the CI were developed in house. A few concrete examples include:
- Research Data Archive services - public interface can be found at: https://rda.ucar.edu/
- Parallel Python tools for post production of NetCDF files and specifically climate data: https://www2.cisl.ucar.edu/tdd/asap/parallel-python-tools-post-processing-climate-data
- System Accounting Manager (SAM) on HPC systems https://www2.cisl.ucar.edu/user-support/systems-accounting-manager (currently NCAR specific)
- VAPOR is the Visualization and Analysis Platform for Ocean, Atmosphere, and Solar Researchers. VAPOR provides an interactive 3D visualization environment that can also produce animations and still frame images https.://www.vapor.ucar.edu/
- NCAR Command Language - NCL is an interpreted language designed specifically for scientific data analysis and visualization.

All tools were primarily developed with the needs of the Atmospheric science community in mind. All components are available for reuse except for SAM. SAM could be customized and utilized by others but would require some generalization or site specific customization.

What external CI capabilities and services and/or externally developed tools (if any) does the facility use and who provides them? How were these tools identified and what criteria was used to select the tools?

A good number of external CI capabilities and/or externally developed tools are in use at NCAR within the Computing and Information Systems Lab (CISL).. Highlights include:
- NCAR Data Sharing Service - Globus Toolkit - https://www.globus.org/
- NCAR also utilizes XDMoD as part of the suite of tools used to manage the HPC resources - http://open.xdmod.org/

Within the NCAR Wyoming supercomputing center two commercial packages are in use to control, manage and monitor the facility.
- The core of the facility utilizes Building Automation, hardware, software and sensors from Johnson Controls Inc. based on the Metasys Building Automation System http://www.johnsoncontrols.com/buildings/building-management/building-automation-systems-bas
- More recently NCAR has deployed an advanced system to allow higher fidelity sampling of

the electrical infrastructure. Those components were provided by Schneider Electric Software LLC. under their Wonderware brand.
These two commercial packages were purchased utilizing a formal RFP process and were evaluated by a technical team, business team and pricing team. Technical requirements were developed in partnership with external engineering firms.

List up to 3 of your most and least favorite CI components with a 1 sentence explanation for each. What aspects about the facility CI and its operation would you like to share as best practices?

The three most used CI components are the High Performance Computing systems, High Performance Disk Storage (GLADE) and the tape archive HPSS. The HPC systems are regularly see greater than 90% system utilization. GLADE similarly has been exceptionally popular providing common shared space across HPC, data analysis and visualization platforms. Finally the HPSS based archive system is still the cornerstone of data archival at NCAR and in some respects is too popular:
- HPC systems utilize test and development hardware that is much smaller scale but provides capabilities to not impact production work while upgrading, patching or adding new tools to the user environment. Once changes to the test environments are stable the teams can then upgrade or change the large HPC environments. Here complexity and scale provide significant challenges.
- The GLADE environment is technically challenging providing a very large (50PB) high performance InfiniBand storage environment. However the technical challenges are only one component of the environment, user retention policies and management of quotas are equally as challenging.
- HPSS presents a more financial challenge. Historical archival storage policies were predicated on computing being expensive but storage being cheap. Currently those economic assumptions are no longer valid and CISL has embarked on modifications to storage policies. That effort is too new but may become a best practice.

What aspects of the facility CI and its operation do you see as challenges/gaps? Are there any pitfalls/mistakes you would like to share? What aspects would you be interested in outsourcing?

We see human capital as possibly one of our most challenging areas currently. Expertise in HPC, large data storage and IT environments are in high demand. We often find recruiting staff a challenge especially where some areas like data analytics and data science are in significant demand in the commercial as well as research sectors. Keeping pace with salaries in a challenging federal environment is proving difficult.
Closer to the facility operation level we are seeing highly dynamic HPC energy consumption based on computing workloads. All HPC vendors are actively pursuing power saving capabilities all the way down to the chip level, turning down clocks or components on demand. Overall this is a good thing as computing systems of the past were notoriously

wasteful. However, computing components that turn up and down on computing timescales (sub seconds) may not be a match for traditional building automation systems or more broadly utility providers. Large changes in electrical demand influence mechanical cooling systems as well as the capacity of the utility. The NWSC has a highly energy efficient design that adapts to the demands of the CI housed in the facility.
Key Risks
Workforce development, recruiting and retention are a significant risk.
What CI-related workforce development activities does your facilities engage in?
NCAR has a number of efforts underway as we see workforce development as critical. The NWSC has been utilized as a teaching laboratory with 7 summer interns over the last 5 years working within the facility. Within that timeframe, 3 women and 2 minority students have been through three-month intensive summer internships. All but two of those students have remained in fields engaged with large CI. CISl also manages the Summer Internships in Parallel Computational Science (SIParCS). The goal of the SIParCS program is to make a long-term, positive impact on the quality and diversity of the workforce needed to use and operate 21st century supercomputers. Graduate students and undergraduate students (who have completed their sophomore year by summer 2017) gain significant hands-on experience in high-performance computing and related fields that use HPC for scientific discovery and modeling. More recently the Operations Manager at the NWSC has been engaged as part of the state of Wyoming Workforce Development Council. Wyoming in particular is looking to develop greater inroads specific to large computing facilities with more traditional trades, community colleges and non-traditional students.
What do you see as your key new CI requirements and challenges in the next 5-10 years
Specific to modeling and simulation we see a highly disruptive CI environment with significant computing architecture diversity on the horizon and new clear winners. Heterogeneous computing architectures are now commonplace but the complexity and scale remain challenging. There is also an explosion of data and data resources that has long been promised but we are starting to see with greater clarity. New methods such as machine learning offer some promise but there are many paths and options. NCAR certainly doesn't have the capability to explore all possible paths and will need to partner across many disciplines to find answers.
Do you have any other suggestions for the workshop?
Not at this time

Affiliation	Name	E-mail
Incorporated Research Institutions for Seismology (IRIS)	Tim Ahern, University of Washington	tim@iris.washington.edu

What percentage of the facility CI was developed in-house versus by reusing existing solutions?
Most components have been developed in house over the 30 years life of the DMC. Of course commercial and open source software systems are used when appropriate such as DBMS software. Much of our infrastructure is somewhat domain specific such as reception of real time data and tools that work with domain specific data.
What external CI capabilities and services and/or externally developed tools (if any) does the facility use and who provides them? How were these tools identified and what criteria was used to select the tools?
We use commercial software for virtualization (VmWare), PostgreSql for DBMS software, commercial geolocation software. All external tools were acquired using IRIS purchasing guidelines, multiple bids etc.
List up to 3 of your most and least favorite CI components with a 1 sentence explanation for each. What aspects about the facility CI and its operation would you like to share as best practices?
1) Web services, methods to abstract time series and metadata access both internally and externally 2) storage RAID indexing scheme to improve access to commodity RAID 3) Synchronization of data versions across multiple storage systems (1 primary and 1 secondary at each of the DMC and the ADC)
What aspects of the facility CI and its operation do you see as challenges/gaps? Are there any pitfalls/mistakes you would like to share? What aspects would you be interested in outsourcing?
Scalability. Access to seismological data can be episodic especially after earthquakes. Also certain preprocessing services can exceed our internal capabilities. The promise of cloud resources has potential but not yet realized.
Key Risks

Loss of key personnel and their knowledge. NSF budgets are making facilities like our more and more vulnerable.
What CI-related workforce development activities does your facilities engage in?
Both NSF and commercially sponsored training courses. We participate as time and financial resources allow
What do you see as your key new CI requirements and challenges in the next 5-10 years
Reducing the cost to maintain our infrastructure and finding external resources perhaps cloud, that can meet our demands and fit our way of doing business not theirs.
Do you have any other suggestions for the workshop?
Nothing at this time, not able to spend much time on this.....

Affiliation	Name	E-mail
UNAVCO	Fran Boler, UNAVCO	fboler@unavco.org

What percentage of the facility CI was developed in-house versus by reusing existing solutions?
Essentially all components of UNAVCO's CI have been developed in house. This includes data handling for data arriving at UNAVCO from multiple varieties field instrumentation and from a variety of providers, archiving, and distribution functions. Most of the CI that aids in data handling is not available for reuse since it is highly customized. An exception is the GNSS preprocessing software tool called "teqc", which is widely shared with the community. Selected CI components have been developed in partnership with other institutions and are shared with them including SAR web services developed via the NASA SSARA project is shared with the Alaska Satellite Facility; and the Geodesy Seamless Archive Centers open source software was developed with NASA ACCESS support by UNAVCO with UCSD and NASA's Crustal Dynamics Data Information Systems. GSAC is widely shared.
What external CI capabilities and services and/or externally developed tools (if any) does the facility use and who provides them? How were these tools identified and what criteria was used to select the tools?
Certain proprietary software provided by sensor manufacturers for handling raw data are part of UNAVCO's CI. These are prescribed when a manufacturer is selected as a sensor provider. Much of UNAVCO's SAR data handling infrastructure is currently being migrated to the XSEDE cloud. Commercial cloud storage is employed as one of our backup strategies.
List up to 3 of your most and least favorite CI components with a 1 sentence explanation for each. What aspects about the facility CI and its operation would you like to share as best practices?
The data systems that we operate (software and hardware) that receive, handle and deliver GNSS data to our external customer base have the largest user base and are used 24/7. We have been "saved" many times over by having failover systems at the ready for the inevitable hiccups in systems.
What aspects of the facility CI and its operation do you see as challenges/gaps? Are there any pitfalls/mistakes you would like to share? What aspects would you be interested in outsourcing?
A gap is lack of adequate resources to keep software and to a lesser extent hardware up to date. Functionality is regularly added through time as new component software systems, and this functionality is developed with technologies reflecting the era during which it was

developed, with some attempt to see into the future; these components tend to remain part of operational infrastructure (we call them legacy components, but they are still key to accomplishing our tasks). All along the way technical debt is incurred, and of course technology moves ahead. This is a further challenge to moving capabilities to the cloud. We are trying to slowly and on a trial basis move components to the cloud. Legacy components are a further risk as it becomes increasingly difficult to find programmers with appropriate skillsets to maintain them. The priority is almost never to rebuild these older systems as long as they continue to operate. Another challenge is the wide variety of technologies in use in the Earth Sciences to meet CI needs of various domains. Trying to cover all bases is nearly impossible; trying to identify which technologies will emerge as most useful is a challenge for all. The EarthCube initiative is clearly exposing/highlighting this.
Key Risks
Key risks are related to the technical debt described in a previous section. Another key risk is looming retirement of staff members with decades of domain knowledge and in-depth knowledge of our CI components. Further, there is strong competition in our geographic area for skilled CI workers.
What CI-related workforce development activities does your facilities engage in?
We send staff members to training. We engage interns.
What do you see as your key new CI requirements and challenges in the next 5-10 years
Making use of the cloud (with appropriate return on investment). Continuing to track and identify trends in technologies and being able to respond nimbly. Managing functionality demands under resource constraints.
Do you have any other suggestions for the workshop?
Not at this time

Affiliation	Name	E-mail
IceCube	Gonzalo Merino, University of Wisconsin Madison	gonzalo.merino@icecube.wisc.edu

What percentage of the facility CI was developed in-house versus by reusing existing solutions?
1) Data management software, handling data archive, transfer from the south pole and replication to long term archives. 2) Software framework to manage distributed workloads. Used to manage and bookkeep all the IceCube simulation production. In both cases, others could use, but this does not happen yet.
What external CI capabilities and services and/or externally developed tools (if any) does the facility use and who provides them? How were these tools identified and what criteria was used to select the tools?
1) South Pole broadband satellites SPTR, DSCS and Skynet. Provided by NASA, through USAP. This is the only available service for daily bulk data transfer from the South Pole. ~100Gbytes/day. 2) Tape storage for long term data archive. Provided by collaborating institutions NERSC and DESY-Zeuthen. These institutions already operate large scale automated tape facilities for several experiments. The service is offered as in-kind contribution to the Collaboration. 3) Open Science Grid. Providing access to millions of CPU hours in opportunistic resources. Also, operating core Grid services that provide us access to IceCube collaborating sites in Europe and Canada. We have been participating in OSG for several years. Distributed computing, and in particular opportunistic computing, represents a big advantage in our field where a lot of the data processing and analysis is pleasantly parallel. 4) XSEDE. Part of the IceCube simulation chain relies on GPUs. We started requesting allocations in GPU-capable XSEDE resources in 2016 to enlarge the computing capacity available for IceCube and increase the analysis potential. 5) Globus data transfer service (globus.org). Convenient data transfer service used to schedule/steer data transfers from UW-Madison to archive locations: NERSC and DESY-Zeuthen. Selected because it provided the needed functionality (integrity, retries, etc) currently at no cost. Also, interested in ongoing developments to interface more efficiently the HPSS tape system at NERSC with Globus (file integrity, performance).
List up to 3 of your most and least favorite CI components with a 1 sentence explanation for each. What aspects about the facility CI and its operation would you like to share as best practices?

1) Main data processing cluster at UW-Madison. Large CPU and GPU cluster coupled to a multi-petabyte filesystem (Lustre) used by ~300 researchers to analyze the IceCube data. The most challenging part to operate is the storage, including monitoring, accounting, etc. However, operating our own Lustre cluster seems to still be the most cost effective solution for our size (~6 Petabytes of disk). 2) User-friendly scalable/elastic computing infrastructure: OSG and HTCondor have provided great capabilities so far in this front. However, we still see a lot of room for improvement in the user experience: higher efficiency, ease of use, interface to cloud resources, etc.
What aspects of the facility CI and its operation do you see as challenges/gaps? Are there any pitfalls/mistakes you would like to share? What aspects would you be interested in outsourcing?
Every time we have been able to leverage existing 3rd party services to build our infrastructure around them, we have seen benefits in doing that. From large archive storage facilities, to data transfer services, to workload management services, our lesson learnt is that it seems worth for us to invest on having a solid interface with existing services rather than trying to replicate them, or reinvent the wheel.
Key Risks
With the use of external services, there comes dependencies and risk. Mitigation strategies are therefore an important topic. In our case, several of these external services are coming from the academic ecosystem, so some coordination inside or between agencies could address part of the risk. Part of it would be ensuring that those common services that many researchers depend on, are sustainable.
What CI-related workforce development activities does your facilities engage in?
Assisting to various workshops and conferences in the field: NSF cyberinfrastructure, Open Science Grid, National Data Service ...
What do you see as your key new CI requirements and challenges in the next 5-10 years
Understanding how to best adapt IceCube analysis code to new emerging computing architectures and software frameworks such as manycore, GPU, FPGA, machine learning and data analytics frameworks, etc and engage the workforce with the required skills that we need to make this happen. Hiring and retaining this personnel is getting increasingly difficult as we compete head-on with the IT private industry.
Do you have any other suggestions for the workshop?
Not at this time

Affiliation	Name	E-mail
NSCL	Andreas Stolz, Michigan State University	stolz@nscl.msu.edu

What percentage of the facility CI was developed in-house versus by reusing existing solutions?
Data acquisition and analysis software framework (NSCLDAQ/SpecTcl/DDAS), available to others. Controls software (EPICS) development, available to others. Business process software; custom and customized applications.
What external CI capabilities and services and/or externally developed tools (if any) does the facility use and who provides them? How were these tools identified and what criteria was used to select the tools?
Data acquisition (DAQ) and experimental data analysis on Linux based infrastructure. Commodity PCs/Servers. Storage using commodity hardware and ZFS/Linux. This is widely used, freely available software and low cost. DAQ is developed in-house. Analysis applications are typical freely available physics applications (GEANT, ROOT, etc.) Business process: ERP (IFS software), Sharepoint workflows and document management. Engineering software? Solidworks etc. Networking/Internet – external access provided by MSU
List up to 3 of your most and least favorite CI components with a 1 sentence explanation for each. What aspects about the facility CI and its operation would you like to share as best practices?
Infrastructure – virtualization: Normal for enterprise infrastructure, but does require expertise for support. Sharepoint: Used for business processes, collaboration etc. Again requiring developer and administrator expertise. Security: Network and systems security including technical controls themselves and the workload around maintaining and documenting same. Adopting configuration management tools and testing deployment processes. System configuration – maintaining stable operations along with ongoing software changes and security updates.
What aspects of the facility CI and its operation do you see as challenges/gaps? Are there any pitfalls/mistakes you would like to share? What aspects would you be interested in

outsourcing?
Security is ongoing challenge.
Key Risks
Main risks are similar to any enterprise: security and disaster recovery.
What CI-related workforce development activities does your facilities engage in?
Participating in relevant workshops. CI Security training for all users.
What do you see as your key new CI requirements and challenges in the next 5-10 years
Providing increased data access to outside visitors and experimenters in face of increasing dataset sizes and security restrictions. Future DAQ systems for FRIB experiments.
Do you have any other suggestions for the workshop?
Not at this time

Affiliation	Name	E-mail
International Ocean Discovery Program (IODP)	Jim Rosser, Texas A&M University	jrosser@tamu.edu

What percentage of the facility CI was developed in-house versus by reusing existing solutions?
Several CI components are developed and maintained in-house: instrument host data uploaders, web services, web science applications, databases, business applications (procurement, inventory, crew tracking). Yes, these are available to others for reuse, but, in most cases, would require extensive effort.
What external CI capabilities and services and/or externally developed tools (if any) does the facility use and who provides them? How were these tools identified and what criteria was used to select the tools?
Our approach is to focus on JRSO core competencies and leverage commodity services from other organizations when possible. For example, Texas A&M University provides many shared services that we use to support JRSO operations, including email; directory services; storage services; web conferencing; video streaming; software training; cloud storage; financial, travel and HR management systems; cybersecurity assessment tools; software procurement; project management assistance, etc.
List up to 3 of your most and least favorite CI components with a 1 sentence explanation for each. What aspects about the facility CI and its operation would you like to share as best practices?
1. WAN (including VSAT) operations and support. Sustaining highly available WAN services is quite challenging when the research vessel (JR) operates globally. 2. Oracle ODAs. Oracle ODAs significantly increased JRSO database engine performance. However, there has been a steep learning curve for configuring and maintaining this capability. 3. Cybersecurity. Minimizing security risk while supporting international customers who bring many different personal devices onboard the JR and expect assured access to the ship's portfolio of science lab services (e.g., LAN, server storage, application and database services).
What aspects of the facility CI and its operation do you see as challenges/gaps? Are there any pitfalls/mistakes you would like to share? What aspects would you be interested in outsourcing?

Minimizing security risk while supporting international customers who bring many different personal devices onboard the JR and expect assured access to the ship's portfolio of science lab services (e.g., LAN, server storage, application and database services).
Key Risks
Commercially available tools are increasingly cloud-based (e.g., Adobe Creative Suite, macOS apps, etc.). Our meager communication bandwidth supporting the JR rules those out. Yet, many software publishers provide no alternative. This issue is probably unique to facilities operating in low bandwidth, high latency environments, and probably also applies to organizations, such as DoD, that operate isolated networks (SIPRNet, JWICS, etc). This is a growing problem that continues to challenge us.
What CI-related workforce development activities does your facilities engage in?
Technology specific training for all aspects of infrastructure, software development and data management.
What do you see as your key new CI requirements and challenges in the next 5-10 years
Better WAN link for the JR. Adoption of automation/configuration management tools, such as Chef, Ansible, Salt, etc. Making data more discoverable.
Do you have any other suggestions for the workshop?
Not at this time

Affiliation	Name	E-mail
CHESS	Werner Sun, Cornell University	wms8@cornell.edu

What percentage of the facility CI was developed in-house versus by reusing existing solutions?
Our high-availability clusters and Compute Farm were developed using commodity hardware and open-source software, assembled and configured in-house to meet the requirements of our facility. These configurations could be shared with other facilities.
What external CI capabilities and services and/or externally developed tools (if any) does the facility use and who provides them? How were these tools identified and what criteria was used to select the tools?
We provide CHESS users with remote data download capabilities using Globus. We selected this tool for its excellent performance and because of its widespread adoption in the NSF Large Facility community.
List up to 3 of your most and least favorite CI components with a 1 sentence explanation for each. What aspects about the facility CI and its operation would you like to share as best practices?
High-availability Linux server clusters form the backbone of our CI. We use them for our central file systems, core infrastructure services, web and database servers, and hardware control systems. In commissioning these clusters, we gained experience with selecting free and open-source software and commodity hardware solutions without sacrificing reliability and performance. The CHESS data acquisition system is a central repository that receives raw data from multiple input streams and provides access for offline analysis and processing. We developed backup, archive, and rotation procedures to ensure disk access to two run-cycles' worth of data and tape retrieval for all previous data.
What aspects of the facility CI and its operation do you see as challenges/gaps? Are there any pitfalls/mistakes you would like to share? What aspects would you be interested in outsourcing?

We would be interested in learning about methods for provisioning temporary accounts and implementing fine-grained authorization for CHESS users.
Key Risks
We face an increasingly challenging cybersecurity threat landscape. We are always seeking ways to balance securing our facility control systems while maintaining usability, access, and productivity.
What CI-related workforce development activities does your facilities engage in?
Online tutorials, managerial and technical trainings.
What do you see as your key new CI requirements and challenges in the next 5-10 years
Upgrades to the scientific capabilities of the CHESS facility will result in increased data throughput and volumes, which will eventually exhaust a single system's ability to both serve as the data store and the access point. We may need multiple ingress and separate analysis systems.
Do you have any other suggestions for the workshop?
Not at this time

Affiliation	Name	E-mail
National Radio Astronomy Observatory (NRAO)	Brian Glendenning, NRAO	bglenden@nrao.edu

What percentage of the facility CI was developed in-house versus by reusing existing solutions?
100% (based on open source software), yes
What external CI capabilities and services and/or externally developed tools (if any) does the facility use and who provides them? How were these tools identified and what criteria was used to select the tools?
Amazon AWS (modest), NSF XSEDE (experimental); Convenience/capability (AWS), cost (XSEDE)
List up to 3 of your most and least favorite CI components with a 1 sentence explanation for each. What aspects about the facility CI and its operation would you like to share as best practices?
1. The CASA data reduction package is a large (2M SLOC) package both used for internal operations use and downloaded by facility users (2k downloads per year). 2. Our "pipelines" embed expert knowledge in a python scripting framework for automated science production. 3. Our computing infrastructure has multiple "archive" storage clusters, with attached Lustre and computational clusters for data processing. We have to take the long view - we have usable data from 40 years ago, our software packages live for decades.
What aspects of the facility CI and its operation do you see as challenges/gaps? Are there any pitfalls/mistakes you would like to share? What aspects would you be interested in outsourcing?
Keeping software packages reasonably high-performance over decades is an issue for us.
Key Risks

Durable agreements with HPC facilities, IaaS research clouds, International compatibility with user authentication mechanisms etc.
What CI-related workforce development activities does your facilities engage in?
Ph.D. student / Post-doc engagement with writing research codes. Summer / co-op students.
What do you see as your key new CI requirements and challenges in the next 5-10 years
See final bullet points in white paper.
Do you have any other suggestions for the workshop?
Not at this time

Affiliation	Name	E-mail
Ocean Networks Canada	Benoit Pirenne	bpirenne@oceannetworks.ca

What percentage of the facility CI was developed in-house versus by reusing existing solutions?
The Oceans 2.0 was entirely developed in house, starting in 2005. The code is not in the public domain owing to the decision made by ONC to pursue commercial applications of the system.
What external CI capabilities and services and/or externally developed tools (if any) does the facility use and who provides them? How were these tools identified and what criteria was used to select the tools?
External tools include standard tools such as OS (Linux), Java, Javascript and attendant libraries; Oracle as an RDMS, Cassandra for non-relational data... ERDDAP was integrated to provide standard access to specific data types. Jira for supporting all aspect of the development, including time sheets and billing on a per project basis Confluence for internal and external documentation
List up to 3 of your most and least favorite CI components with a 1 sentence explanation for each. What aspects about the facility CI and its operation would you like to share as best practices?
Until recently, the challenging elements included: - Cassandra: performance issues with the tool and the complexity of the fine-tuning required , Java memory allocation issues, difficulty with profiling complex code to understand where memory and time are actually spent, despite having an advanced test environment
What aspects of the facility CI and its operation do you see as challenges/gaps? Are there any pitfalls/mistakes you would like to share? What aspects would you be interested in outsourcing?
Continuously evolving the technology and the services available and getting the continued funding for the required manpower. Providing easy to use data discovery interfaces that will be addressing user needs in the face of growing instrumentation, observing locations and expanding time
Key Risks
Risks include-maintaining the level of funding to enable continuous improvements to the

facility: a CI is never over! Mitigation requires making management and funding agencies understand that.
What CI-related workforce development activities does your facilities engage in?
We have had large fractions of the team of 20+ software engineers attend classes in: - the Agile Scrum methodology - usability - Kaisen
What do you see as your key new CI requirements and challenges in the next 5-10 years
- As the facility continues to grow, a continuous emphasis on verification of our scalability, and possible adaptation will be necessary. - The support of multiple clients, re-organizing into a multi-project based entity - Need to support critical customers (e..g, Public Safety) with defined SLAs
Do you have any other suggestions for the workshop?
Not at this time

Affiliation	Name	E-mail
Oregon State University, College of Earth, Ocean, and Atmospheric Sciences, Regional Class Research Vessel Program	Christopher Romsos	cromsos@coas.oregonstate.edu

What percentage of the facility CI was developed in-house versus by reusing existing solutions?

The most significant CI component built in-house is our "datapresence" system. In a nutshell, the datapresence system captures and archives data from resident (or visiting) sensors, replicates the information shoreside, and presents the information to both the shipboard and shoreside science parties for use/consumption. The datapresence system includes functionality for data quality assessment, flagging, alert and user notification.

Other CI components developed in-house include several databases for project management including a risk-register database application.

Yes, these components are available for others to use.

What external CI capabilities and services and/or externally developed tools (if any) does the facility use and who provides them? How were these tools identified and what criteria was used to select the tools?

There is a high likelihood that the most if not all RCRVs shall be provisioned with satellite service through HiSeasNet at UCSD (https://hiseasnet.ucsd.edu/), though some UNOLS ships are experimenting with going out and negotiating their own contracts for satellite service opting (out of the HighSeasNet program in areas where better deals can be struck such as the Gulf of Mexico).

We, the RCRV datapresence developers, are currently formalizing an MOU with Leidos Antarctic Support contractors to share components of our acquisition and visualization code. Part of this process includes choosing an open source license under which to distribute software.

Lastly, we've incorporated data and map services (hosted locally aboard the ship) from the Marine Geoscience Datasystem at Lamont-Doherty Earth Observatory (LDEO) into our real-

time displays for scientific situational awareness. Specifically, the Global Multi-Resolution Topography Data Synthesis provides our base layer for the map interface http://www.marine-geo.org/portals/gmrt/ Other sources of thematic background information for this interface are provided by NOAA Fisheries, Office of Coast Survey, USGS, and various academic sources.

List up to 3 of your most and least favorite CI components with a 1 sentence explanation for each. What aspects about the facility CI and its operation would you like to share as best practices?

1) Ship to shore (and back) data replication over high latency, low bandwitdh satellite networks. This problem, akin to the Long Fat Network problem of high bandwidth-delay product, is the most challenging issue that we are working on. We've had good success in increasing our throughput by optimizing the TCP window and buffer sizes and are now looking at managed WAN optimizatoin solutions to provide this service.
2) Cybersecurity is another challenge for the project. The RCRVs shall be equipped with integrated monitoring control systems to cover everything from bridge to engine room systems. Securing these online systems is a priority and a challenge.

What aspects of the facility CI and its operation do you see as challenges/gaps? Are there any pitfalls/mistakes you would like to share? What aspects would you be interested in outsourcing?

At this project phase (construction) we don't yet have lessons learned to share.

Key Risks

Key risks include security and expertise. As indicated the RCRVs shall present a significant CI advancement from current. To mitigate each of these risks we have an operations plan that includes support and oversight (budget and personnel) from a Class Management Office. However, the level of expertise for the technical support personnel (Marine Technicians) that sail with the ships will have to rise. Evidence to support this expertise risk can be gleaned from organizations that have recently taken operations responsibility for new research vessels.

What CI-related workforce development activities does your facilities engage in?

Ah, a perfect follow-up question. A key component of our operations plan during transition to operations and post-delivery under Class Management will be technology transfer and

training for new operators. We expect much of this initial ' workforce development' to take the form of hands on work during transition but additional training will be made possible through the Class Management Office during operations. In addition to periodic training we have staff that shall travel to each vessel on a rotating schedule (multiple visits per year) to inspect sensor systems, perform calibrations and maintenance, as well as conduct specific training while on a site visit.
What do you see as your key new CI requirements and challenges in the next 5-10 years
BYOD IoT sensors - We must keep abreast of security and integration issues these devices present. On-Prem IaaS and PaaS - These industry trends or options are attractive but difficult to implement under the current model of support and operations (see expertise risk above). Cybersecurity - Particularly as it applies to on-board integrated monitoring and control systems.
Do you have any other suggestions for the workshop?
Not at this time

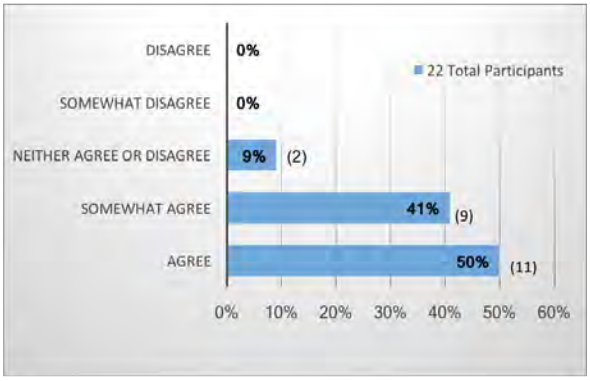
Appendix E. Post-workshop Survey Responses

Pre-workshop Survey Questionnaire

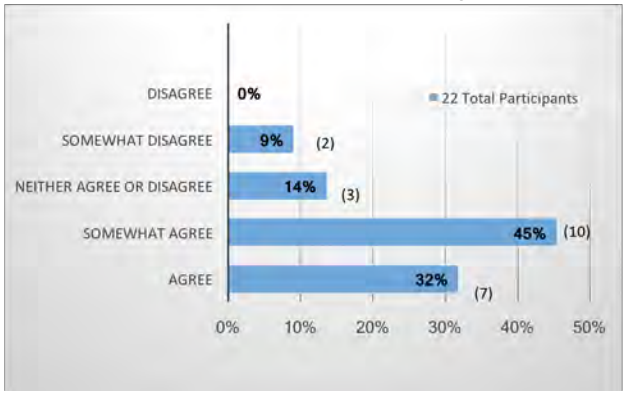
1. Contact Information
    - Name
    - Facility/project represented
    - Role at facility/project
    - Email
2. Rate the workshop (1-5) on its ability to meet each of the goals listed below:
    - Understand current CI architecture and operations best practices at the large facilities.
    - Identify common requirements and solutions, as well as CI elements that can be shared across facilities.
    - Enable CI developers to most effectively target CI needs and gaps of large facilities.
    - Explore opportunities for interoperability between the large facilities and the science they enable.
    - Develop guidelines, mechanisms, and processes that can assist future large facilities in constructing and sustaining their CI.
    - Explore mechanisms and forums for evolving and sustaining the conversation and activities initiated at the workshop.
3. As follow-up activities, centered on facilities CI, would you like to see (select all that apply):
    - Smaller and more focused workshops on specific technical topics?
    - A common portal with information about facilities
    - A common portal with information about CI
    - A discussion forum
    - Community calls/seminars
    - Community training opportunities
    - Other
4. Should this workshop be held again? If so, what should be the focus?
5. Would you like to contribute a short science success highlight to our website?

Rate the workshop (1-5) on its ability to meet each of the goals listed below

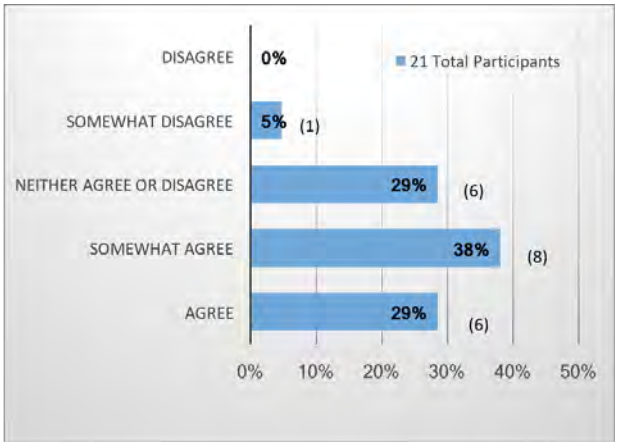
Workshop provided an understanding of current CI architecture &operations best practices at the facilities.



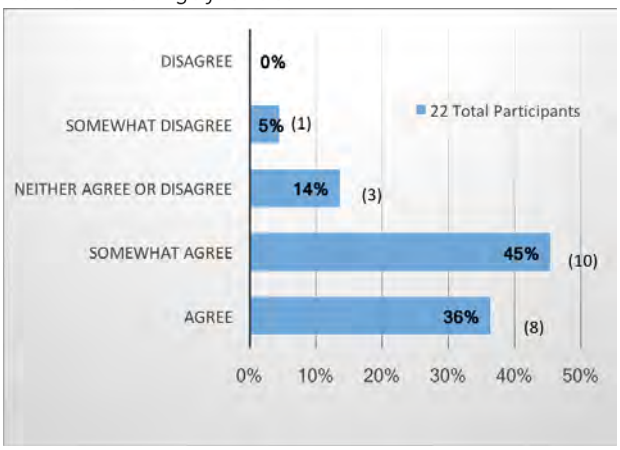
Workshop identified common requirements and solutions, as well as CI elements that can be shared across facilities.



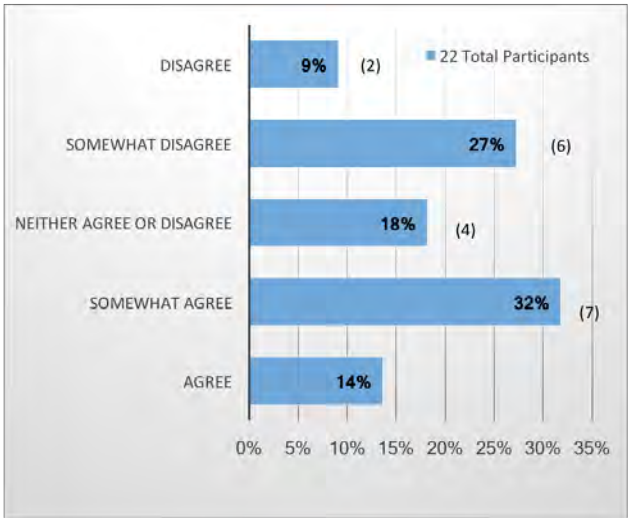
Workshop enabled CI developers to most effectively target CI needs and gaps of large facilities.



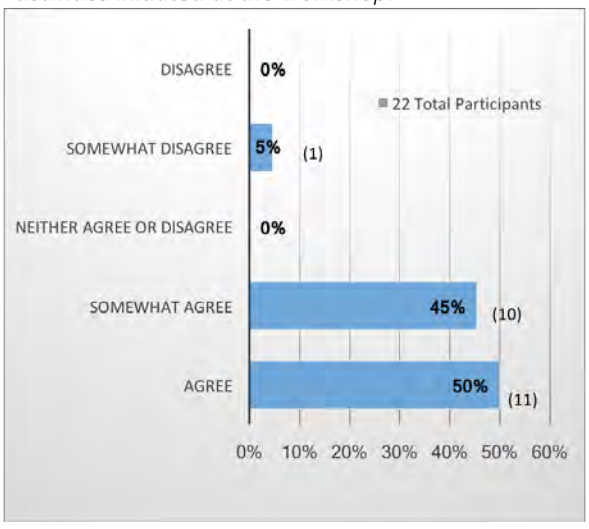
Workshop explored opportunities for interoperability between the large facilities.



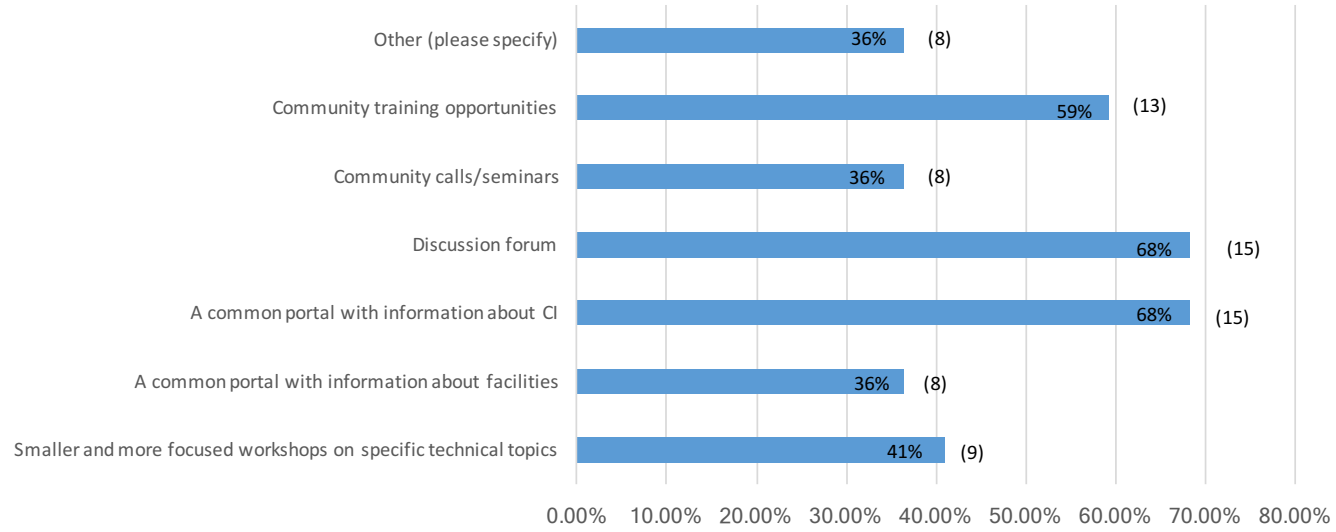
The workshop developed guidelines, mechanisms, and processes that can assist future large facilities in constructing and sustaining their CI.



Workshop explored mechanisms and forums for evolving and sustaining the conversation and activities initiated at the workshop.



As follow-up activities, centered on facilities CI, would you like to see (select all that apply):



Responses under the "Other" category were:

- Communication channels to build community relationships and identity.
- CI effectiveness metrics.
- It would be great to determine ways to engage those who were in the audience more, by having small working group discussions. The smaller sessions seemed to get more of the people involved in the discussions.
- List of NSF identified technical experts available for consult to discuss implementation of best practices.
- A standing, cross-facility working group with small amount of funding, a funded center of excellence for CI for large facilities, similar to CTSC is for cybersecurity.
- For me this was most useful because I got to talk with multiple facilities that I am either already working with, or just starting to work with. Repeating this kind of meeting annually in fall is useful.
- Up to date entries on CI in the large facilities manual.
- CI management within Higher Education; dealing with the user end of the spectrum and recognizing the accountability of that build.

Should this workshop be held again? What should focus be?

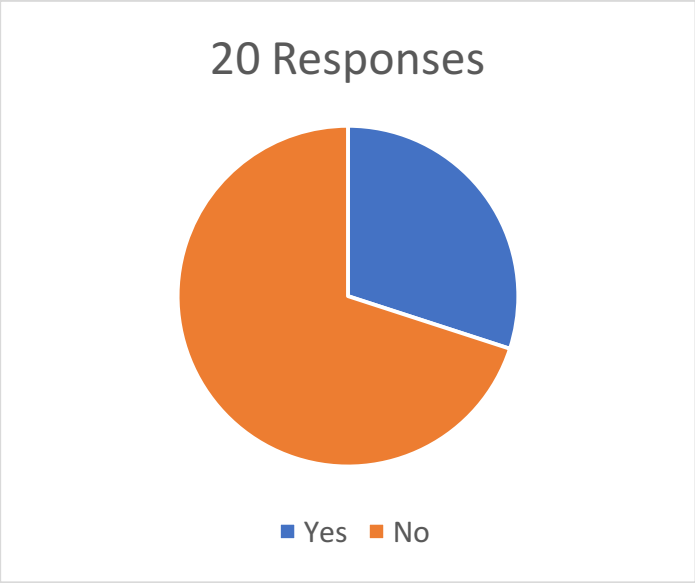
There were 20 responses, which are listed below.

1. A follow up workshop that is more focused on collaboration opportunities, interoperability of resources across facilities, and areas for shared cyberinfrastructure, training, etc. would be valuable. A joint workshop with EarthCube's Council of Data Facilities is worth considering.
2. Yes. Finding specific areas of CI overlap among facilities and forming partnerships with the goal of sharing/leveraging efforts for greater efficiency.
3. Yes. Share best practices for Large Facilities CI.
4. I personally found the workshop to be of great value in understanding the NSF LF/CI landscape. Polling the community prior to the workshop to identify focus and goals enables participants to have active role in the discussions and drive community progress. I would suggest repeating this activity next year.
5. Best practices to fuse CI functions with scientists served by them.
6. 1) Large scale Big Data, Big Compute research - tools and algorithms 2) Invite students to the event 3) A session from user perspectives on how unique users are using CI 4) a session on management best practices for finding and retaining talent to filling pipeline.
7. Yes. I suggest adding more time to have facilities present their specific technical solution to generic services (DevOps, Authentication, Asset Management, Software Config Management, etc). Covering pros/cons. Also, how can the NSF be more proactive to support intra-LSF collaboration? Develop communication channels, or form an organization to lead guidance on CI (as CTSC does for CyberSecurity)?
8. Yes. I think the focus should be on working to provide a sustainable home for cyberinfrastructure resources including repositories of best practices, registries of available shared services, human resources, etc.
9. Yes. Continued dialog on challenges and solutions from sites.
10. Yes! The evolution of XSEDE/SPs in support of Large Facilities Science.
11. Yes, but only with a charter to take the results of the first two workshops into action, otherwise this is just discussion that leads nowhere.
12. Yes. Annually in fall works great!
13. Yes, sort of. I'd suggest a general workshop like this one every (say) five years, to allow for the changing landscape of CI and large facilities. But annually I think it would be good to focus a bit more - best practices in particular areas (with input from the LFs on what those areas should be). I don't think we really got very deep into best practices in this one, there wasn't really time.
14. I am not sure that there was much consensus on how to move forward. It is a difficult topic and not an easy thing to do. I think developing some forums for technical people to discuss common approaches and best practices would be more fruitful.
15. Yes. Focus on developing a mechanism to increase communication and collaboration between LF CI folks. I want to be able to seek help from this group of experts when I'm planning new capability and solving problems.
16. Hold again if discernible impact observed.
17. I think the workshop was important as a step towards creating a community for NSF Large Facilities CI. However, I think it is just the start of the conversation. I answered "somewhat agree" to most of the questions above, because there simply was not enough time to delve

into solutions. I think an in-person workshop should be held again, to help enable this nascent community, to ultimately lead to solving challenges of interoperability and sustaining CI.

18. I found the workshop worthwhile. The panel format seemed about right. Some speakers did a better job than others in providing general overviews that were relevant to facilities other than their own.
19. I would attend the workshop again. Sometimes results aren't immediately obvious, but that does not mean that NSF should not have some role in facilitating bridges across centers. Doing nothing would certainly not have a better outcome. For this past workshop I got the feeling that many of the presentations focused on specific CI implementation details pertaining to individual centers. Further, sometimes the break-out sessions were monopolized by one or two participants with strong opinions based on their own experiences. Although it's interesting to see what other centers are doing, many of their implementation details could not be easily translated into re-usable tools and/or general Best Practices. I believe this is a cross-cutting issue across centers, because of contention between results-driven science and process-driven engineering of CI. For instance, software can be built in short order to deliver a science result (e.g. a paper), but the way the software was written, validated, documented and packaged determines if anyone will ever be able to re-use all or part of that software. The latter portion might add to the cost in the short term, and delay the science result (e.g. Pick two: Good, Fast, Cheap). In any case, it's good that we're talking about "Best Practices" in CI but converging on them is not easy and requires a long-term sustained investment in dollars and manpower that is not easily guaranteed by cyclic funding.
20. Yes, a similar focus is great, and maybe add some centered discussion from the point of view of CI users.

Would you like to contribute a short science success highlight to our website?



Contact Information, Facility Represented, and Role at Facility of Respondents

Mohan Ramamurthy	Unidata Data Services and Tools for Geoscience	Director
Fran Boler	UNAVCO	Data archive manager
Andreas Stolz	National Superconducting Cyclotron Laboratory	CCF Operations Department Head
James Marseteller	CTSC CCoE (PSC)	Chief Information Security Officer
Tom Gulbransen	Battelle - NEON	Cyberinfrastructure Data Products Development
Forough Ghahramani	Rutgers Discovery Informatics Institute	Associate Director
Eric Cross	DKIST	IT Manager - NSO
Werner Sun	CHESS	IT Director
Philip Gates	IODP	IT Support Supervisor
Pamela Hill	NCAR	Manager of storage systems group / Storage system architect
David Halstead	NRAO	CIO / Head of IT
Jeff Kantor	LSST	Senior Manager Project Office
fkw@ucsd.edu	OSG	Executive Director
andy adamson	Gemini Observatory	Associate Director, Operations
Tim Ahern	IRIS Data Services	Director of Data Services
Jim Rosser	IODP	IT Director
Brian Glendenning	NRAO	Data Management and Software management
Stace Beaulieu	Woods Hole Oceanographic Institution (WHOI)	Coordinator of WHOI's Ocean Informatics initiative
Albert Lazzarini	LIGO Laboratory	Deputy Director
Sean McManus	NOAO	Head of Data Management Operations
Rafael Ferreira da Silva	Pegasus Workflow Management System	Computer Scientist
Caroline McHugh	Rutgers University	Coordinator RDI2

Appendix F. White Papers

White papers were expected to address the following:

- A brief description of the facility, its science mission, and the community (including size, makeup – number of individuals, number of institutions, etc.) and a URL for more information.
- A description of the key products/services of the facility (data, software, services, etc.).
- A brief description (including a figure) of the facility CI (e.g., its architecture, key services/components, underlying infrastructure), how it is deployed/distributed, and its operation. What is the median age since deployment of the key CI components?

LIGO LABORATORY
MS 100-36
PASADENA CA 91125

TEL: 626.395.2129
FAX: 626.304.9834

MEMORANDUM

DATE: June 20, 2017

SUBJECT: **LIGO white paper for the NSF Large Facilities Cyberinfrastructure Workshop**

- ***A brief description of the facility, its science mission, and the community (including size, make up – number of individual, number of institutions, etc.). Please include a URL for more information.***

The Laser Interferometer Gravitational-wave Observatory (LIGO) comprises a distributed NSF facility with two 4 km x 4 km interferometers, separated by a baseline for 3,002 km, located on the DOE Hanford Nuclear Reservation north of Richland, WA and north of Livingston, LA. LIGO Laboratory is operated jointly by the California Institute of Technology and the Massachusetts Institute of Technology for the NSF under a cooperative agreement with Caltech and MIT as a sub-awardee. LIGO also includes major research facilities on the Caltech and MIT campuses.

The two gravitational wave detectors are operated in coincidence. LIGO detected gravitational waves from the inspiral and merger of a binary black hole system on 14 September 2015, heralding the opening of a new observational window on the Universe using gravitational waves to detect and study the most violent events in the cosmos.

LIGO serves the worldwide gravitational wave community through the LIGO Scientific Collaboration, consisting of over 40 institutions in 15 countries. This international collaboration comprises about 1,100 members. LIGO also has MOUs covering joint operations with the EU Virgo Collaboration and the Japanese KAGRA Collaboration.

More information about LIGO may be obtained at the following URL:
https://www.ligo.caltech.edu

- ***A description of the key products/services of the facility (data, software, services, etc.)?***

The key data product generated by LIGO is a time series recording relative changes in length between the two 4km arms of each LIGO interferometer. These strain measurements (~3 TByte/y) record audio-frequency perturbations in the local space-time metric at each Observatory at the level of 1 part in 10^{22}. This is the primary observable from the LIGO experiment, recording the signature of gravitational waves

passing through each detector. To inform data analysis efforts searching the strain data for gravitational waves, and to understand and improve the performance of the LIGO instruments, an additional ~200k channels of environmental monitors and internal instrument channels are recorded (1.5 PByte/y). The strain data are distributed in low-latency (seconds) to computing clusters running analysis pipelines to generate gravitational-wave triggers for external Astronomical observations for transient events on a timescale of 1 minute. The bulk data are locally archived at each LIGO Observatory and distributed over the Internet to a central data archive on a timescale of 30 minutes. The central data archive currently holds 7 PByte of LIGO observations in perpetuity.

LIGO data analysis software is released using native Linux packaging (.rpm and .deb) and pre-installed on dedicated computing resources via standard Linux software repositories. For computing on shared resources the software is distributed via the CERN Virtual Machine Filesystem (CVMFS) and containerized with Docker, Singularity, or Shifter. Similarly, the key science data are pre-staged on dedicated computing resources ahead of analysis, and distributed to shared computing resources via CVMFS or GridFTP as needed by computing tasks. Metadata that describe LIGO observations and candidate signals from data analysis are stored in databases with custom tools for ingestion and querying.

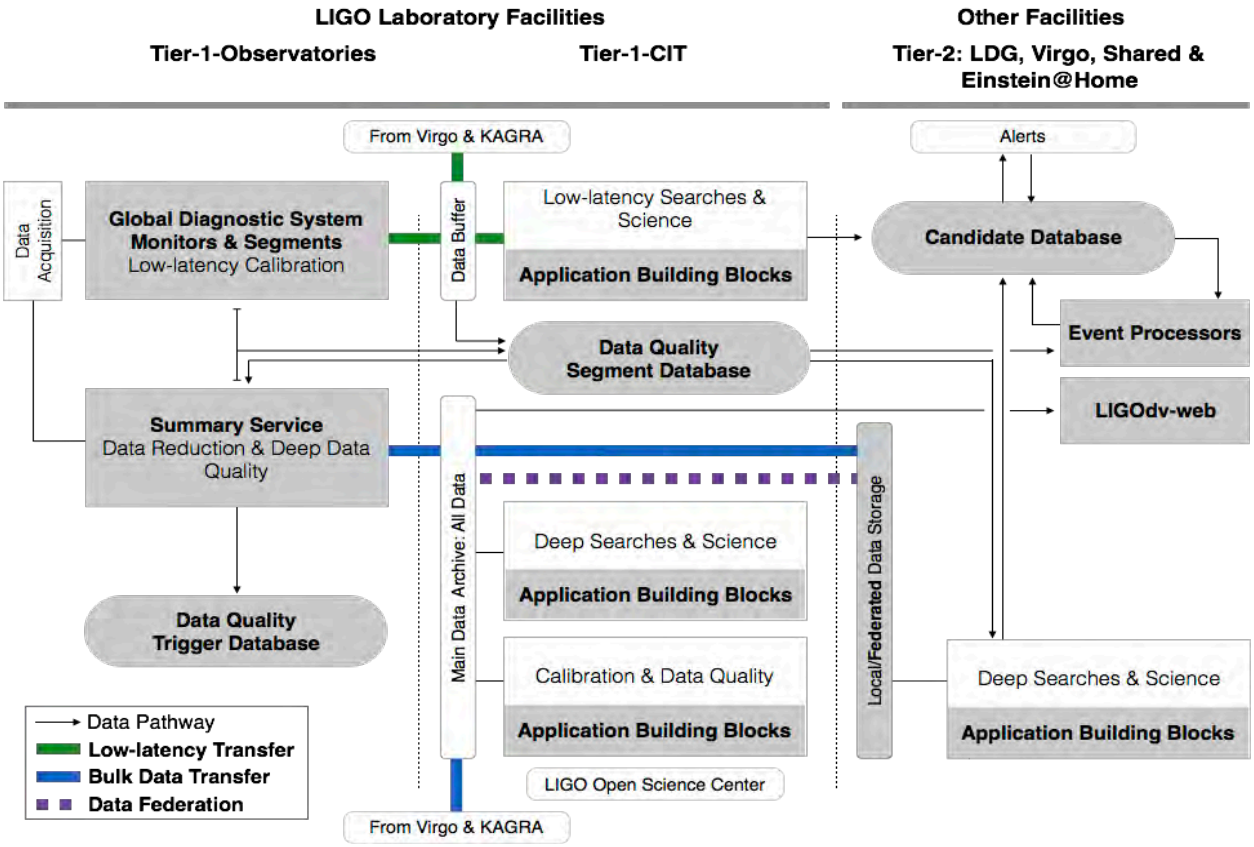
- ***A brief description (including a figure) of the facility CI (e.g., its architecture, key services/components, underlying infrastructure), how it is deployed/distributed, and its operation. What is the median age since deployment of the key CI components.***

LIGO data analysis computing overwhelmingly consists of embarrassingly parallel workflows executed on high-throughput (HTC) resources. The majority of LIGO computing is provided by internal LIGO Scientific Collaboration (LSC)-managed clusters, but a growing fraction is provided by external shared resources. These resources are integrated into LIGO's computing environment via the Open Science Grid, and consist of a variety of dedicated and opportunistic campus, regional, and national clusters, Virgo scientific collaboration resources, and XSEDE allocations.

LIGO relies on HTCondor for its internal job scheduling, and uses both DAGMan and the Pegasus WMS for large-scale workflow management on top of HTCondor. In addition, LIGO uses the BOINC infrastructure to manage its single largest data analysis task (the search for continuous wave signals) via Einstein@Home running on volunteer computers as a screen saver. For Single Sign-On and other Identity

and Access Management functions, LIGO relies on Shibboleth, Grouper, InCommon, and CILogon. The underlying authentication infrastructure is built on Kerberos and authorization information if reflected in LDAP.

For distributed data management, LIGO relies on CVMFS, StashCache/Xrootd, Globus GridFTP, and a variety of in-house CI tools and services to complement and integrate these tools.



Unidata (http://www.unidata.ucar.edu) is a community data facility for the atmospheric and related sciences, established in 1984 by U.S. universities with sponsorship from the National Science Foundation (NSF). The Unidata Program Center (UPC), the program office for Unidata and the nexus of activities related to Unidata's mission, is managed by the University Corporation for Atmospheric Research (UCAR), a consortium of over 109 member universities and academic affiliates providing science in service to society.

Unidata exists to engage and serve researchers and educators dedicated to advancing the frontiers of Earth System science. The program's aim is to help transform the conduct of research and education in atmospheric and related sciences by providing well-integrated, end-to-end data services and tools that address many aspects of the scientific data lifecycle, from locating and retrieving useful data, through the process of analyzing and visualizing data either locally or remotely, to curating and sharing the results.

Specifically, the UPC:

- Acquires, distributes, and provides remote access to real-time meteorological data.
- Develops software for accessing, managing, analyzing, visualizing, and effectively using geoscience data.
- Provides comprehensive training and support to users of its products and services.
- In partnership with others, facilitates the advancement of tools, standards and conventions.
- Provides leadership in cyberinfrastructure and fosters adoption of new tools and techniques.
- Assesses and responds to community needs, fostering community interaction and engagement to promote sharing of data, tools, and ideas.
- Advocates on behalf of the community on data matters, negotiating data and software agreements.
- Grants equipment awards to universities to enable and enhance participation in Unidata.

Unidata is governed by its community. Representatives from universities populate standing and *ad hoc* committees that set policies for the program, provide first-hand feedback from users of program software and services, and offer guidance on individual projects

While Unidata's primary mission of serving universities engaged in atmospheric science education and research has remained unchanged through the years, the evolution and broad usefulness of its products and services have greatly enlarged its initial user base. Today, the Unidata community includes users from all sectors in over 200 countries, including nearly 2500 academic institutions and more than 80 research labs. Simultaneously, Unidata's activities and responsibilities have also grown as community needs have evolved. Despite the growth in users and enhanced scope of its activities, according to a 2010 survey conducted by the Unidata Users Committee, 97% of the respondents indicated that they were either satisfied or highly satisfied with Unidata's overall service to the community.

In the following sections we highlight some key quantitative and qualitative metrics that are used to gauge Unidata's success. These indicators offer a peek at Unidata's impact and how its cyberinfrastructure plays an irreplaceable role in advancing research, education, and outreach goals of its community. It should be noted that the UPC provides many of these metrics to its governing committees as part of its regular status reports.

Data services

Delivery of geoscience data to universities in near real time via the IDD system is at the core of Unidata's mission and is extremely

Table 1: Growth of IDD data volume	2008	2013	2016
Volume of data pushed to IDD sites	2.7 TB/day	13 TB/day	18.8 TB/day

important to our university community. Table 1 gives an idea of the explosive growth in the volume of data delivered via the IDD over the past decade.

While the IDD uses a "push" mechanism to deliver data automatically as it becomes available, Unidata's remote data access mechanisms (including THREDDS Data Servers, ADDE servers, RAMADDA servers, and EDEX servers) also provide roughly **670 GB/day** to community members.

Software and support

Unidata community members rely on the UPC to provide access to a variety of software packages for data transport, management, analysis, and visualization. Table 2 shows how many community members have downloaded the software packages that the UPC develops and supports over the past five years.

In addition to providing the software for download, UPC developers also provide the community with direct technical support via electronic mail. The support system is heavily used, with **more than 21000 support queries handled by UPC staff in the past five years**.

Table 2: Software Package Downloads	2012-2016
AWIPS	7700
GEMPAK	8400
IDV	43600
LDM	10100
McIDAS	300
netCDF-C Libraries*	566800
netCDF-Java Libraries (Common Data Model)	41900
TDS	9000
UDUNITS	22700

Appendix: A description of the key products/services

Data Distribution

The UPC coordinates the Internet Data Distribution system (IDD), in which hundreds of universities cooperate to disseminate near real-time earth observations via the Internet. While the "push" data services provided by the IDD system are the backbone of Unidata's data distribution services, the UPC also provides on-demand "pull" data services via THREDDS, ADDE, and RAMADDA data servers.
The UPC's data servers are not classified as "operational" resources, but they nonetheless have a 99.96% uptime record and are used heavily by educational sites that lack the resources to store IDD-provided data locally, or to operate their own data servers (see **Error! Reference source not found.**). UPC's servers are housed in a UCAR co-location computer facility for reliability, and share UCAR's Internet2/National Lambda Rail connectivity, which provides access to ample bandwidth for Unidata's needs.

Software

A variety of software packages are developed, maintained, and supported by the UPC:

NetCDF

Unidata's netCDF (network Common Data Form) is a freely distributed collection of data access libraries that provide a machine-independent data format that is self-describing, portable, scalable, appendable, sharable, and archivable – all important qualities for those who wish to create, access, and share array-oriented scientific data. NetCDF permits easy access to array-based, multi-dimensional datasets, a task that can be difficult when using other common storage schemes. NetCDF has been adopted widely by the atmospheric sciences community, and is especially popular among climate and ocean modelers. For example, model output datasets for the Sixth Assessment Report of the Intergovernmental Panel on Climate Change must be submitted in netCDF format, using the associated Climate and Forecast (CF) metadata conventions. The resulting large base of netCDF users and data has led to support for the format in more than 80 open source packages and many commercial applications including MATLAB and IDL.

Common Data Model & THREDDS Data Server

Unidata's Common Data Model (CDM) provides an interface for reading and writing files in netCDF and a variety of other scientific data formats. The CDM uses metadata to provide a high-level interface to geoscience-specific features of datasets, including geolocation and data subsetting in coordinate space. Unidata's THREDDS Data Server (TDS) builds on the CDM to allow for browsing and accessing collections of scientific data via electronic networks. Data published on a TDS are accessible through a variety of remote data access protocols including OPeNDAP, OGC Web Map Service (WMS) and Web Coverage Service (WCS), NetCDF Subset Service (NCSS), and HTTP.

Integrated Data Viewer

Unidata's Integrated Data Viewer (IDV) is a 3D geoscience visualization and analysis tool that gives users the ability to view and analyze a rich set of geoscience data in an integrated fashion. The IDV brings together the ability to display and analyze satellite imagery, gridded data (such as numerical weather prediction model output), surface observations (METARs), upper air soundings, NWS NEXRAD Level II and Level III RADAR data, NOAA National Profiler Network data, and GIS data, all within a unified interface. The IDV integrates tightly with common scientific data servers (including Unidata's TDS) to provide easy access to many real-time and archive datasets. It also provides collaborative features that enable users to easily share their own data holdings and analysis products with others.

AWIPS II & GEMPAK

AWIPS II is a weather forecasting, display, and analysis package currently being developed by the NWS and NCEP. Because many university meteorology programs are eager to use the same tools used by NWS

forecasters, Unidata community interest in AWIPS II is high. UPC staff have worked closely with NCEP staff during AWIPS II development in order to devise a way to make it available to the university community.

NCEP has stated that GEMPAK applications will be migrated from GEMPAK/NAWIPS into AWIPS II for the National Centers. The UPC will likewise facilitate a migration from GEMPAK/NAWIPS to AWIPS II for the university community.

Rosetta

The Rosetta project at the UPC is an effort to improve the quality and accessibility of observational data sets collected via datalogging equipment. The initial goal of Rosetta is to transform unstructured ASCII data files of the type commonly generated by datalogging equipment into the netCDF format, while minimizing disruption to existing scientific workflows.

Local Data Manager

The Unidata Local Data Manager (LDM) system includes network client and server programs designed for event-driven data distribution. It is the fundamental component of the IDD system. The LDM is used by hundreds of sites worldwide, and is integrated into the National Weather Service's AWIPS II package.

McIDAS

The Man-computer Interactive Data Access System (McIDAS) is a large, research-quality suite of applications used for decoding, analyzing, and displaying meteorological data. The older McIDAS-X system, developed by the University of Wisconsin's Space Science Engineering Center and supported by Unidata, is gradually being replaced by the IDV and by McIDAS-V (which is based on the IDV).

UDUNITS

Unidata's UDUNITS supports conversion of unit specifications between formatted and binary forms, arithmetic manipulation of units, and conversion of values between compatible scales of measurement.

RAMADDA

The Repository for Archiving, Managing and Accessing Diverse Data (RAMADDA) is a vibrant and growing technology initially developed by Unidata and now managed and developed as an open source project. Unidata integrates RAMADDA functionality into the IDV, provides training and support, and contributes code to the project. In addition, Unidata makes extensive use of RAMADDA to support community and collaborative projects, and actively facilitates its deployment in the university community.

2-Dimensional Crystal Consortium - Materials Innovation Platform (2DCC-MIP)

Facility Description:

2DCC Vision

Advance discovery-driven research into the growth, properties and applications of 2D chalcogenide crystals for next-generation electronics through the development of state-of-the-art synthesis and characterization tools within a multidisciplinary user environment to enable expansive national leadership in this important area.

2DCC Mission

1. Accelerate discovery in 2D chalcogenide materials by operating a world-class user facility that includes:
    a) a closed loop iterative collaboration of thin film and bulk growth synthesis techniques, *in situ* characterization, and predictive modeling of growth mechanisms and processes
    b) a community of practitioners that combines the expertise of an in-house research program and external users
    c) open sharing of knowledge, best practices, and publication-quality data
2. Provide access to synthesis, in-situ characterization and theory/simulation user facilities including instrumentation and expertise to users through a competitive proposal process
3. Maintain a vibrant in-house research program in synthesis, characterization and theory/simulation of 2D chalcogenides to drive advances in the field
4. Engage a diverse user base from academia, government and industry in the U.S. and internationally and increase participation of women and minorities underrepresented in science and technology through diverse representation in staffing and research activities.

Key products/services:

The 2DCC platform is defined by three major components: In-house research, user facility, and education/outreach in support of the research mission

Science Drivers (In-house research) -- The 2DCC research priorities are organized by four science drivers that are motivated by the unique properties of layered materials that often emerge in ultrathin or few-layer films, necessitating atomic-level control of film growth mode, stoichiometry, point defects and structural imperfections. The science drivers are: Physics of 2D Systems, Epitaxy of 2D Chalcogenides, Next-generation 2D Electronics, and Advanced Characterization and Modeling.

User Facility – The user program focuses on three main facility components:
(1) *Synthesis and In situ Characterization of Thin Films*
(2) *Bulk Crystal Growth*
(3) *Theory/Simulation*

The user program is focused on the synthesis of 2D chalcogenides for next generation electronics and includes priorities that are accomplished by a community of practitioners that collaborate among the in-house research and external user programs. Over time,

priorities will be adjusted by meritorious peer-reviewed proposals, user committee recommendations, and input from the 2DCC external advisory committee.

Education/Outreach – The 2DCC offers programs that address *engagement of a diverse user base* from academia, government and industry in the U.S. and internationally and *broadening participation* of women and minorities underrepresented in STEM. Education/Outreach programs include: 1) an education series that includes executive course, tutorials and hands-on training; 2) a monthly webinar series that is broadcasted live; 3) major sponsorship and participation in the annual Graphene and Beyond workshop; 4) a travel extension program for 2DCC faculty to visit PUIs and MSIs and present the work of the 2DCC and highlight opportunities for involvement; and 5) Opportunities for summer extended stays for users wishing to spend intensive time training at the facility.

Facility CI:
Theory efforts in the 2DCC-MIP aims at accurately modeling the growth of two-dimensional chalcogenides with multiscale methods and simulating a broad range of characterization techniques from first-principles, both in deep collaboration with 2DCC experimentalists. As a user facility focused on synthesis, the 2DCC does not have a dedicated CI; computational work is divided between two facilities, with the majority of the current workload managed by the Penn State Institute for CyberScience Advanced CyberInfrastructure (ICS-ACI), and future works supported by XSEDE research allocation on the Louisiana State University superMIC (420k CPU hours).

The physical infrastructure of ICS-ACI located in the Penn State University Park campus, where about 50% of the facility's power and equipment resources are dedicated to supporting the infrastructure. The ICS-ACI cluster consists of over 1200 nodes on Linux 6, with high-performance Ethernet or Infiniband interconnects. The queueing system supports interactive and batch jobs. In addition, a Guaranteed-Response Time (GReaT) model is offered, guaranteeing queue times of at most one hour to participating subscribers (2DCC users included). In the current phase the 2DCC theory team accesses 60 256-GB nodes and 10 TB shared storage under an allocation of 1000k CPU hours released on a quarterly basis, with expansion planned for the next phase. Software required by the 2DCC team are provided in the ICS-ACI cluster software stack, including highly parallel quantum chemistry and molecular dynamics codes, along with software libraries that allow for custom compilation. The median age of key CI components is less than 1 year.

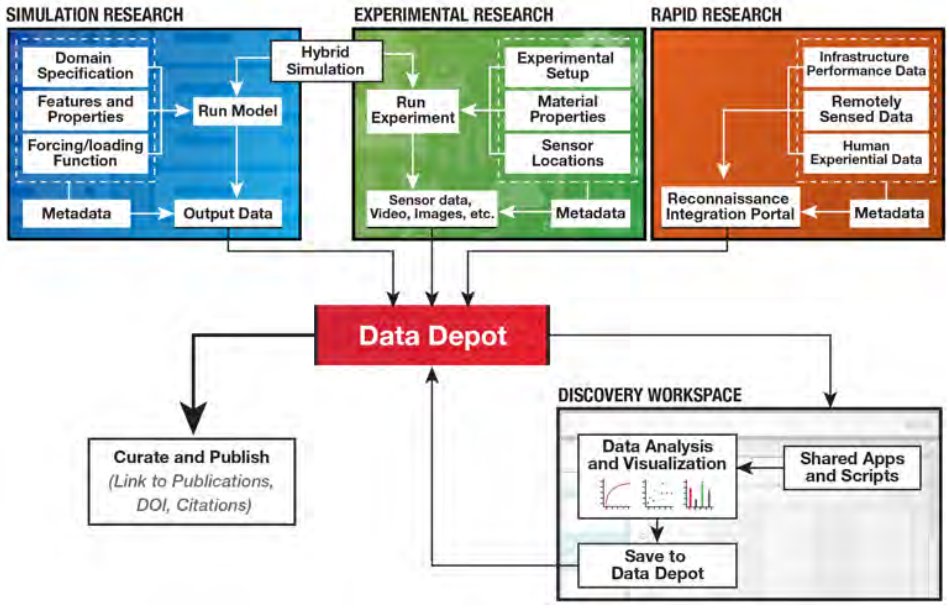
DesignSafe – Cyberinfrastructure for NSF Natural Hazards Engineering Research Infrastructure
https://www.designsafe-ci.org/

Natural hazards engineering plays an important role in minimizing the effects of natural hazards on society through the design of resilient and sustainable infrastructure. The DesignSafe cyberinfrastructure has been developed to enable and facilitate transformative research in natural hazards engineering, which necessarily spans across multiple disciplines and can take advantage of advancements in computation, experimentation, and data analysis. DesignSafe allows researchers to more effectively share and find data using cloud services, perform numerical simulations using high performance computing, and integrate diverse datasets such that researchers can make discoveries that were previously unattainable. This white paper describes the design principles used in the cyberinfrastructure development process, introduces the main components of the DesignSafe cyberinfrastructure, and illustrates the architecture of the DesignSafe cyberinfrastructure.

A cyberinfrastructure is a comprehensive environment for experimental, theoretical, and computational engineering and science, providing a place not only to steward data from its creation through archive, but also a workspace in which to understand, analyze, collaborate and publish that data. Our vision is for DesignSafe to be an integral part of research and discovery, providing researchers access to cloud-based tools that support their work to analyze, visualize, and integrate diverse data types. DesignSafe builds on the core strengths of the previously developed NEEShub cyberinfrastructure for the earthquake engineering community, which includes a central data repository containing years of experimental data. DesignSafe preserves and provides access to the existing content from NEEShub and adds additional capabilities to build a comprehensive CI for engineering discovery and innovation across natural hazards. DesignSafe has been developed along the following principles:

Create a flexible CI that can grow and change. DesignSafe is extensible, with the ability to adapt to new analysis methods, new data types, and new workflows over time. The CI is built using a modular approach that allows integration of new community or user supplied tools and allows the CI to grow and change as the disciplines grow and change.

Provide support for the full data/research lifecycle. DesignSafe is not solely a repository for sharing experimental data, but is a comprehensive environment for experimental, simulation, and field data, from data creation to archive, with full support for cloud-based data analysis, collaboration, and curation in between. Additionally, it is the role of a cyberinfrastructure to continue to link curated data, data products, and workflows during the post-publication phase to allow for research reproducibility and future comparison and revision.



Provide an enhanced user interface. DesignSafe supplies a comprehensive range of user interfaces that provide a workspace for engineering discovery. Different interface views that serve audiences from beginning students to computational experts allow DesignSafe to move beyond being a "data portal" to become a true research environment.

Embrace simulation. Experimental data management is a critical need and vital function of the CI, but simulation also plays an essential role in modern engineering and must be supported. Through DesignSafe, existing simulation codes, as well as new codes developed by the community and SimCenter, are available to be invoked directly within the CI interface, with the resulting data products entered into the repository along with experimental and field data and accessible by the same analytics, visualization, and collaboration tools.

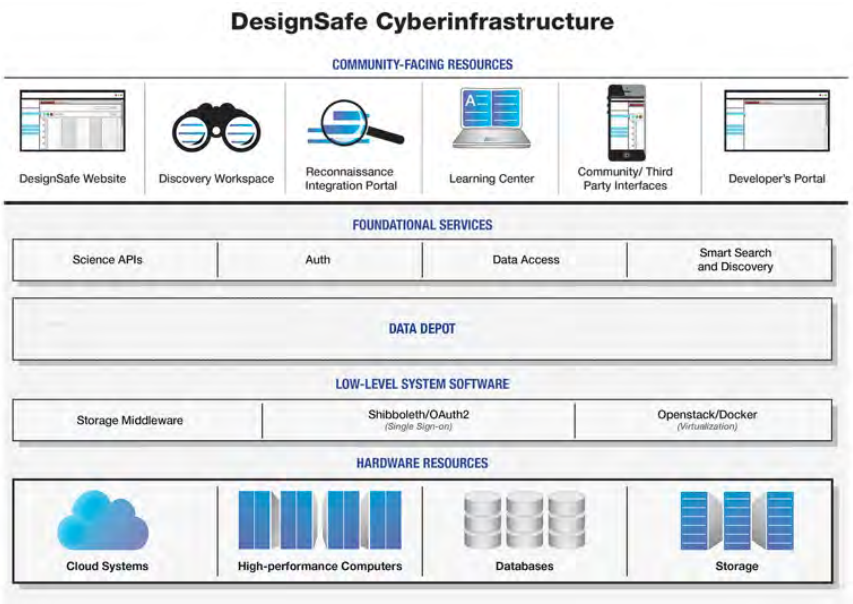
Provide a venue for internet-scale collaborative science. As both digital data captured from experiments and the resolution of simulations grow, the amount of data that must be stored, analyzed and manipulated by the modern engineer

is rapidly scaling beyond the capabilities of desktop computers. DesignSafe embraces a cloud strategy for the big data generated in natural hazards engineering, with all data, simulation, and analysis taking place on the server-side resources of the CI, accessible and viewable from the desktop but without the limits of the desktop and costly, slow data transfers. **Develop skills for the cyber-enabled workforce in natural hazards engineering**. Computational skills are increasingly critical to the modern engineer, yet a degree in computer science should not be a prerequisite for using the CI. Different interfaces lower the barriers to HPC by exposing the CI's functionality to users of all skill levels, and best of breed technologies are used to deliver online learning throughout the CI to build computational skills in users as they encounter needs for deeper learning.

The DesignSafe infrastructure provides a comprehensive environment for experimental, theoretical, and computational engineering and science, providing a place not only to steward data from its creation through archive, but also the workspace in which to understand, analyze, collaborate and publish that data. The CI can be described in terms of the services it provides or in terms of the technical components that enable those services.

DesignSafe is architected to comprise the following services and components as shown in the figure:

- **DesignSafe** front end web portal
- The **Data Depot**, a multi-purpose data repository for experimental, simulation, and field data that uses a flexible data model applicable to diverse and large data sets and is accessible from other DesignSafe components. The Data Depot includes an intelligent search capability that allows dynamic creation of catalogs of the held data in an easily understandable way, and that can search ill-structured data with poor or incomplete metadata.
- A **Reconnaissance Integration Portal** that facilitates sharing of reconnaissance data within a geospatial framework.



- A web-based **Discovery Workspace** that represents a flexible, extensible environment for data access, analysis, and visualization.
- A **Learning Center** that provides training and online access to tutorials.
- A **Developer's Portal** that provides a venue for power users to extend the Discovery Workspace or Reconnaissance Integration Portal, and to develop their own applications to take advantage of the DesignSafe infrastructure's capabilities.
- A foundation of **storage and compute** systems at the Texas Advanced Computing Center (TACC), to provide both on-demand computing and access to scalable computing resources.
- A **middleware layer** to expose the capabilities of the CI to developers, and to enable construction of diverse web and mobile interfaces to data products and analysis capabilities
- A marketplace of **Community Defined Interfaces;** the extension capability of the CI allows other projects to leverage DesignSafe to build an interface of their own choosing.

The CI development was initiated in July 2015 upon receiving the NSF award, and was first deployed May 2016. As of June 2017 we have more than 1,100 registered users spanning dozens of institutions around the world.

Cyberinfrastructure at the National Nanotechnology Coordinated Infrastructure (NNCI) Site at Georgia Institute of Technology

The National Nanotechnology Coordinated Infrastructure (NNCI) is an NSF-funded program comprised of 16 sites, located in 17 states and involving 29 universities and other partners. This national network provides researchers from academia, government, and industry with access to university user facilities with leading edge fabrication and characterization tools, instrumentation and expertise within all disciplines of nanoscale science, engineering, and technology. Research undertaken within NNCI facilities is incredibly broad, with applications in electronics, materials, biomedicine, energy, geosciences, environmental sciences, consumer products, and many more. The toolsets of sites are designed to accommodate explorations that span the continuum from materials and processes through devices and systems. There are micro/nanofabrication tools, used in cleanroom environments, as well as extensive characterization capabilities to provide resources for both top-down and bottom-up approaches to nanoscale science and engineering. For more information about NNCI, please visit www.nnci.net. Georgia Tech serves as the coordinating office for the NNCI.

Modeling and simulation play a key role in enhancing nanoscale fabrication and characterization as they guide experimental research, reduce the required number of trial and error iterations, and enable more in-depth interpretations of the characterization results. Various NNCI sites provide a diverse set of software and hardware resources and capabilities. Some of these resources are only available to internal users and some to academic users and some to all interested parties. The rest of this white paper describes the rational behind a major cyberinfrastructure at Georgia Tech and its features and capabilities. This computing resource currently serves only students and faculty at Georgia Tech and is not available for external users.

Science and engineering research is the key to understanding everything in our universe and the best way we can improve the human condition. We are on the cusp of answering fundamental questions in the physical sciences, life sciences, social sciences, and mathematical and computational sciences. As our understanding deepens, we can leverage our basic fundamental knowledge to develop innovative and creative technologies that help drive solutions to the most pressing global problems all enabled by advances in cyberinfrastructure.

Investment in heterogeneous, sustainable, scalable, secure, and compliant cyberinfrastructure is critical to enable future discoveries. Significant resources are needed to address the storage, network bandwidth, and massive computational power required for simulation and modeling across multiple scales. Data-centric computing is also vital, necessitating high-throughput analysis and mining of massive datasets, as well as the ongoing demand for low cost, long-term, reliable storage. Sustained investment in cybersecurity will support sharing of datasets along with greater multi-institution and multi-disciplinary research collaboration. A significant investment in software engineering will enable researchers to leverage the promise offered by public-private, multi-cloud based cyberinfrastructure and emerging new architectures. Some of the greatest risks are an inability to meet workforce demand and the lack of a sustainable funding model. Addressing these issues includes maximizing the steady pipeline of students entering science and engineering careers; creating professional retooling programs; building specialized local and regional teams; and leveraging a range of investment sources including federal, state, municipal and local entities, as well as public-private partnerships (e.g. academic and industry, government and corporate).

Future breakthroughs are reliant on continued investment of national level resources in the path to exascale systems. That said, there are real limitations in an approach that primarily relies on "big iron" systems. More broadly, the perception is a general lack of resources to accommodate large simulations due to smaller jobs that require high-throughput computing. This problem is not likely to be addressed by reaching exascale capacity as there is essentially unbounded demand yet natural boundaries to scalability at many levels. Few researchers have access to funding to port code to new architecture introduced by these “big iron” systems. The national scale resources are also not well suited for small to medium-sized jobs and local institutional support is uneven and inconsistent.

Our existing cyberinfrastructure is also limiting for researchers who need more data-centric systems. Many modern computational tasks are "embarrassingly parallel" and have strong scalability, but available computer clusters and HPC systems are not designed or optimized for such HTC workloads. Examples include data analytics and deep learning workloads. We must develop new systems that can more efficiently support data intensive applications. There are promising technologies for this including modern memory hierarchies, GPUs, and other heterogeneous environments.

In 2009, Georgia Tech created a technology model for central hosting of computing resources that would be capable of supporting multiple science disciplines with shared resources, private resources, and a group of expert support personnel, in support of campus research community. This project is called "Partnership for an Advanced Computing Environment (PACE)." Since its inception, PACE has acquired more than 50,000 cores of high performance computing capability and more than 8 Petabytes of total storage used by approximately 3000 (1500 active) faculty and graduate students. This project provides power, cooling, and high-density racks, as well as a three tiered storage system including home directory, project space, and high transfer rate scratch space across the whole system. On top of storage, compute capabilities are provided both as private resources for a researcher or research group, or as a public resource with access open to researchers on campus through a proposal process for requesting compute cycles. PACE is funded through a mix of central and faculty funding that has proven sustainable is expected to continue with increased growth into the future (Figure 1). Due to this rapid growth, more hosting capability is being planned.

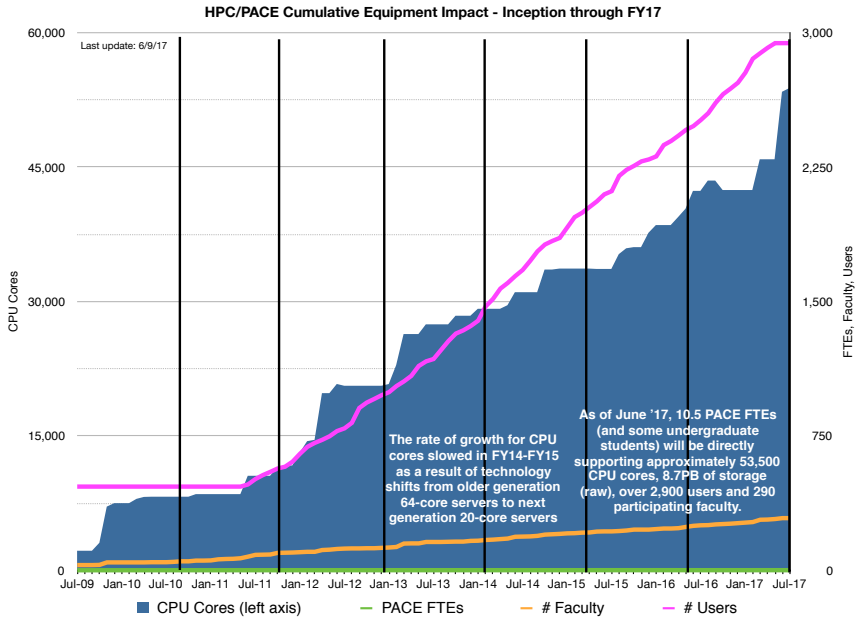


Figure 1. Growth at PACE, represented in cumulative terms of CPU Cores in blue, users in pink, faculty in orange, and PACE full-time employees (FTEs) in green.

A significant investment in software engineering will enable researchers to leverage the promise offered by public-private, multi-cloud based cyberinfrastructure and emerging new architectures. Some of the greatest risks are an inability to meet workforce demand and the lack of a sustainable funding model. Addressing these issues includes maximizing the steady pipeline of students entering science and engineering careers; creating professional retooling programs; building specialized local and regional teams; and leveraging a range of investment sources including federal, state, municipal and local entities, as well as public-private partnerships (e.g. academic and industry, government and corporate).

NRAO Whitepaper Submission to the NSF Large Facilities Cyberinfrastructure Workshop

B.E. Glendenning, Assistant Director, NRAO Data Management and Software Department

1 NRAO TELESCOPES

The National Radio Astronomy Observatory (NRAO, https://public.nrao.edu/) operates the Karl G. Jansky Very Large Array (VLA) near Socorro New Mexico, and is the operating partner (Executive) for the North American part of the Atacama Large Millimeter/Submillimeter Array (ALMA), which operates at a high site near San Pedro, Chile.

Both telescopes are very general purpose. Telescope time is allocated based on a peer-review process from many sub-fields of astronomy. Hundreds of PI groups per year get data, and in addition once the proprietary period has expired (usually one year), the data may be used by other groups for Archival research.

Both telescopes are radio interferometers, which operate by coherently combining the signals of the relocatable antennas (27 for the VLA, 66 for ALMA) in complex central electronics (notably the correlators, which are approximately 0.1 Exa-Op very parallel special purpose supercomputers) which produces raw data, essentially a noisy (electronics, radio-frequency-interference, atmospheric and other environmental effects) irregularly sampled spatial Fourier transform of sky "stacked" over separate frequency channels for up to 4 polarizations.

The electronics are capable of sustaining 1 (VLA) and 16 (ALMA) Gigabytes per second of raw data output, although the data rates are usually averaged down (in time, and frequency) to a small fraction of that (typically 25 Megabytes/second for the VLA, and 6 MB/s for ALMA). This averaging is done both to reduce the computing that is needed, and because many times the science application do not need high data rates. However there are some classes of science observations that are not made because computing capacity is not available.

The raw data is turned into regularly gridded 2-4 dimensional images (axes: position on the sky, frequency or Doppler velocity, polarization) using multi-million line of code software systems produced by the NRAO and our partners. These images (currently: Giga-pixel, coming Tera-pixel, Possible: Peta-pixel) are then typically processed through analysis codes (both produced by NRAO and the wider community) to enable the science to be extracted from the data.

2 CURRENT NRAO COMPUTING PARADIGM

The raw science data from each telescope is buffered at the telescope site (to allow for network outages and periods of high data rate observing), from which it is transferred and ingested into the master archive (in Santiago in the case of ALMA, Socorro NM in the case of the VLA). In the case of ALMA the data is then replicated from the master archive to the "regional" archives, which for North America resides at Charlottesville Virginia. Through an archive search web interface the raw data may be

downloaded by operations staff and the PI group that proposed the observations (after QA in the case of ALMA). The raw data may be freely downloaded by anyone after the (typically) 1-year proprietary period has expired.

After the raw-data for the entire project has arrived in the archive (this could take several different observing sessions), "pipelines" are executed which automatically make derived data products, currently flagged and calibrated raw data for both telescopes, and reference images for the case of ALMA. After some QA is performed, these data products may be downloaded by the PI groups, or by anyone after the proprietary period has expired. NRAO has initiated a "Science Ready Data Products" (SRDP) project to improve the quality of the automatically generated data products, with a goals that: the images should be directly usable for science, to improve the user interfaces, and to allow a human to be in the loop to optimize via high-level guidance the derived data products to be well suited for use in answering particular science questions.

At the moment, almost all VLA derived data products, and many ALMA ones, which are used for the actual science analysis are produced through the manual (including ad-hoc Python scripting) execution of programs from suites of data processing, analysis, and visualization tasks produced by the NRAO. These programs are developed by the NRAO with significant contributions from our ALMA partners, and total about 3M SLOC. This software is available under an open source license, although the NRAO generates executables for common Linux variants and recent versions of MacOS.

The software is executed at a combination of NRAO and user facilities. Our software is downloaded several thousand times per year for use by users (laptops through small clusters). In addition the NRAO allows our users to use our in-house computing facilities through a reservation system. Although our resources are relatively modest (150 16-core compute nodes, 2 PB of fast Lustre filesystem with Inifiniband interconnects), they are well tuned to our software stack, have fast access to the raw data archives, and we allow them to be used interactively (we also have batch queues). That is, they are convenient to use and very suitable for modest problem sizes. Our computing resources are used by a few hundred PI groups per year.

We have experimented with commercial cloud providers (AWS) and national supercomputing centers (XSEDE), but have not made extensive use of either yet, nor have our users.

Key CI improvements areas we would identify are:

- In-the-cloud Elastic, Interoperable, Data Center accessibility
- Machine learning applications (vs. ad-hoc expert knowledge capture in scripts)
- Software sustainability infrastructure
- Visualization and information extraction from multi-peta-pixel multi-dimensional image data

We look forward for the opportunity to discuss these topics at the workshop.

CHESS Facility and Cyberinfrastructure Overview

Devin Bougie and Werner M. Sun
Cornell University, Ithaca, New York 14853, USA

June 19, 2017

Abstract

We present an overview of the Cornell High Energy Synchrotron Source (CHESS), focusing on the facility and key cyberinfrastructure components.

1 CHESS Facility

The Cornell High Energy Synchrotron Source (CHESS) is a NSF-funded National User Facility located on the Cornell University campus in Ithaca, New York. The mission of CHESS is to provide a national hard x-ray synchrotron radiation facility for individual investigators, on a competitive, peer reviewed, proposal basis. With 11 experimental stations, the facility is used by approximately 1,100 investigators per year from over 150 academic, industrial, government, non-profit, and international institutions. CHESS impacts a wide range of disciplines, serving researchers from the physical, biological, engineering, and life sciences, as well as cultural specialists such as anthropologists and art historians. CHESS users conduct studies encompassing, but not limited to, the atomic and nanoscale structure, properties, operando, and time-resolved behavior of electronic, structural, polymeric and biological materials, protein and virus crystallography, environmental science, radiography of solids and fluids, and micro-elemental analysis, and other technologies for x-ray science.

The CHESS facility is hosted by the Cornell Laboratory for Accelerator-based Sciences and Education (CLASSE), which also operates the Cornell Electron Storage Ring (CESR) as the x-ray source for CHESS. Computing services for CHESS are provided centrally by the CLASSE-IT department. The primary computing services used by CHESS are:

- high-speed data acquisition for x-ray detectors at the CHESS experimental stations
- access to and long-term storage of x-ray data collected by CHESS users
- software libraries and parallel computation resources for CHESS staff and users.

More information about the CHESS facility may be found at http://www.chess.cornell.edu.

2 CHESS Cyberinfrastructure

The CLASSE cyberinfrastructure (CI) consists of an interconnected series of high-availability server clusters (HACs), data acquisition systems, control systems, compute farms, and workstations. Most of these systems run either Scientific Linux or Windows on commodity 64-bit Intel-based hardware and are centrally managed using Puppet. The median age of key CI components is approximately 5 years, with an average refresh rate of once every 10 years. The CLASSE CI components most relevant to CHESS are described below and are shown in Figure 1.

2.1 Central Infrastructure

The central Linux infrastructure cluster runs the core CLASSE infrastructure services, including name services, file systems, databases, and web services. Recently, a dedicated oVIrt cluster has been commissioned to run centrally-provisioned virtual machines. These clusters utilize shared 10Gb iSCSI storage domains, and they provide file systems and other basic services to the rest of the lab.

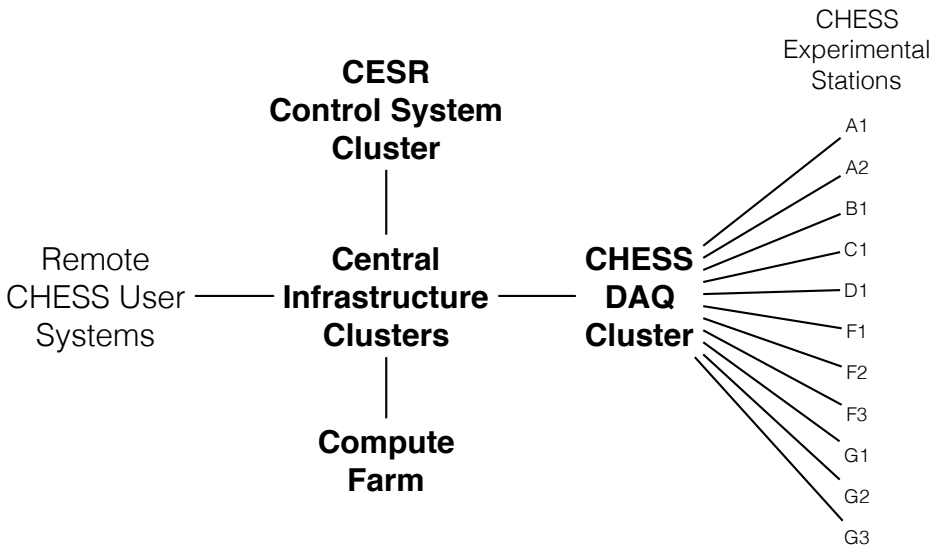


Figure 1: Main components of the CHESS cyberinfrastructure.

2.2 CHESS Data Acquisition (DAQ)

The CHESS data acquisition system runs on a dedicated HAC and provides 10Gb network connections to each experimental station. Data collected at the stations are written directly to the data acquisition system over either NFS or Samba, where it can then be processed on the CLASSE Compute Farm or end-user workstations. CHESS users can also download their data remotely using a Globus server endpoint or via SFTP.

2.3 Compute Farm

The CLASSE Compute Farm is a central resource consisting of approximately 60 enterprise-class Linux nodes (with around 400 cores) with a front-end queueing system that distributes jobs across the Compute Farm nodes. This queueing system supports interactive, batch, parallel, and GPU jobs, and it ensures equal access to the Compute Farm for all users.

2.4 CESR Control System

The CESR control system, responsible for running the particle accelerator that produces x-rays for CHESS, consists of a dedicated Linux HAC. Although the CESR, CLASSE, and CHESS DAQ clusters are essentially identical, the CESR cluster runs many more control system services and is able to operate independently from the CLASSE central infrastructure. This isolation ensures continuity of CESR operations in the event of a power failure or general network outage.

2.5 User Connectivity

Based on their requirements, CHESS users are either granted restricted "external" CLASSE accounts (providing access to station computers and remote access to data) or full CLASSE accounts (providing access to the CLASSE Compute Farm and full interactive desktops, both local and remote).

While collecting data at the experimental stations, CHESS users generally connect their instruments and experimental equipment to a private subnet that is selectively firewalled from the rest of the CLASSE infrastructure. If users require direct write access to the CHESS DAQ filesystems, they may use dedicated station and kiosk computers located at the experimental stations and in other restricted-access locations. Outside the experimental stations, CHESS user data is made available for read-only access through the CLASSE public network.

Research Vessels: Seagoing Datacenters

Jon C. Meyer, Information Systems Manager
UC San Diego, Scripps Institution of Oceanography

Introduction

Scripps Institution of Oceanography (SIO) is a graduate school of UC San Diego and is a world leader in oceanographic field research. SIO supports the operation and/or scientific research of 3 research vessels, a research platform, and is in the primary role in a multi-institution partnership that works with the US Coast Guard to conduct arctic oceanographic research. SIO also manages a cost-saving satellite-based Internet project for research vessels at sea, serving the network-based needs of the majority of University-National Oceanographic Laboratory System (UNOLS) participants.

Key Products/Services

The Ship Operations & Marine Technical Support (SOMTS) department within SIO offers basic and specialized services.

Our most basic (and obvious) service is that of functional and fully equipped seagoing platforms for oceangoing research. These platforms range from a regional research vessel (R/V Robert Gordon Sproul), and Ocean Class research vessel (R/V Sally Ride -- America's newest research vessel) and a Global Class vessel (R/V Roger Revelle -- our flagship). We also support a specialized platform, R/P Flip, which is a platform designed to stably study ocean currents by inverting itself 90 degrees in the water. All platforms come equipped with instrumentation and information systems to acquire commonly useful information about the environment: from seawater temperature and salinity, ocean floor, ocean currents, wind and weather, etc. These systems often operate with other devices as a system of systems, providing cohesive information about a vessel's movement in order to better understand the environment around that vessel.

We also support a number of specialized projects: repeat hydrography, arctic research aboard the USCGC Healy (in partnership with other academic institutions and the US Coast Guard), and support a multi-dish satellite earth station through the HiSeasNet project which has provided affordable Internet to the UNOLS community for the better part of a decade.

Finally, we are in the process of exploring the data delivery mechanism(s) upon completion of scientific missions. At present, data is delivered via “sneakernet” to a data archive/curation project, but as Internet connectivity improves, standardized realtime delivery of data from oceanographic ships at sea should to. Further, modern instrumentation data needs are growing. Newer vessels are installing instruments that produce 100 times more data than other systems; a cohesive, modern data management plan is being sought for these standalone environments.

Deployment

We are in the process of upgrading SIO’s mobile platforms to datacenter-grade computing to provide the redundancy, resiliency and the graceful degradation of equipment that only a no-single-point-of-failure system can provide. Despite redundancies, severe weather and rough seas can make off-ship communication difficult at times; as such a ship needs to be somewhat self-contained when communications go awry.

Working oceanographic equipment (along with the attached computing systems) tend to have a slow upgrade path. Many ships work the majority of the year; an idle ship is expensive. As such, equipment upgrades and maintenance have to be targeted to be as non-disruptive as possible. As such, we are constantly seeking opportunities to proactively deploy and maintain equipment. That said, some of the equipment on-hand does not have clear upgrade paths and it is not rare to find a 10+ year old computer system aboard a ship. Getting such systems to behave reliably can be a losing battle.

Internet connectivity at sea remains challenging to engineer consistently and keep ships online. After a decade of successes, HiSeasNet is looking to the future to re-equip all of UNOLS with modern, maintained satellite communications. Older installations in the fleet are showing signs of wear, and proactivity is needed to keep the fleet communicating well.

Summary

Oceanographic field research is fraught with challenges of being both self-sufficient where it matters, available via network in locations with little infrastructure. SIO is looking to meet these challenges with 21st century solutions, and help lead the charge to produce excellent data from its seagoing research that will be useful and have impact for many years.

National Science Foundation Ocean Observatories Initiative (OOI) Cyberinfrastructure White Paper

Ivan Rodero and Manish Parashar
Rutgers Discovery Informatics Institute (RDI²), {irodero, parashar}@rutgers.edu

INTRODUCTION

The NSF Ocean Observatories Initiative (OOI) is a networked ocean research observatory with arrays of instrumented water column moorings and buoys, profilers, gliders and autonomous underwater vehicles within different open ocean and coastal regions. OOI infrastructure also includes a cabled array of instrumented seafloor platforms and water column moorings on the Juan de Fuca tectonic plate. This networked system of instruments, moored and mobile platforms, and arrays will provide ocean scientists, educators and the public the means to collect sustained, time-series data sets that will enable examination of complex, interlinked physical, chemical, biological, and geological processes operating throughout the coastal regions and open ocean.

The seven arrays built and deployed during construction support the core set of OOI multidisciplinary scientific instruments that are integrated into a networked software system that will process, distribute, and store all acquired data. The OOI has been built with an expectation of operation for 25 years. This unprecedented and diverse data flow is coming from 89 platforms carrying over 830 instruments which provide over 100,000 scientific and engineering data products.

The OOI is funded by the National Science Foundation and is managed and coordinated by the OOI Program Office at the Consortium for Ocean Leadership (COL). Implementing organizations, subcontractors to COL, are responsible for construction and development of the different components of the program. Woods Hole Oceanographic Institution (WHOI) is responsible for the Coastal Pioneer Array and the four Global Arrays, including all associated vehicles. Oregon State University (OSU) is responsible for the Coastal Endurance Array. The University of Washington (UW) is responsible for cabled seafloor systems and moorings. Rutgers, The State University of New Jersey, is implementing the Cyberinfrastructure (CI) component. The OOI data evaluation and education and public engagement team is co-located with the Cyberinfrastructure group at Rutgers University.

OOI CYBER-INFRASTRUCTURE SERVICES

The primary functions of the OOI CI are data acquisition/collection, storage, processing and delivery. The overall architecture of the OOI CI network is shown in Figure 1.

(a) Data Collection and Transmission to the OOI CI: Data is gathered by both cabled and un-cabled (wireless) instruments located across multiple research stations in the Pacific and Atlantic oceans. Once acquired, the raw data (consisting mostly of tables of raw instrument values – counts, volts, etc.) are transmitted to one of three operations centers: Pacific City, directly connected via fiber optic cable to all cabled instruments in the Cabled Array; OSU, an Operational Management Center (OMC) responsible for all un-cabled instrument data on the Pacific coast; and WHOI, the OMC for Atlantic coast-based un-cabled instrument data. The data from the operations centers is transferred to the OOI CI for processing, storage and dissemination.

(b) Data Management, Storage, and Processing: Two primary CI centers operated by the Rutgers Discovery Informatics Institute (RDI²) are dedicated to OOI data management: the West Coast CI in Portland, OR, and the East Coast CI, at Rutgers University. While data from the Cabled Array components are initially received at the Shore Station in Washington, it is the East Coast CI that houses the primary computing servers, data storage and backup, and front-facing CI portal access point, all of which are then mirrored to the West Coast CI over a high-bandwidth Internet2 network link provisioned by MAGPI (Mid-Atlantic GigaPOP in Philadelphia) on the east coast and PNWGP (Pacific-Northwest GigaPOP) on the west coast. The data stores at the OMCs at OSU and WHOI are continuously synchronized with the data repositories located at the East and West Coast CI sites.

(c) Data Safety & Integrity: Data safety and protection is ensured in two ways: data security and data integrity. Data security is addressed through the use of a robust and resilient network architecture that employs redundant, highly available next-generation firewalls along with secure virtual private networks. Data integrity is managed through a robust and resilient information life-cycle management architecture.

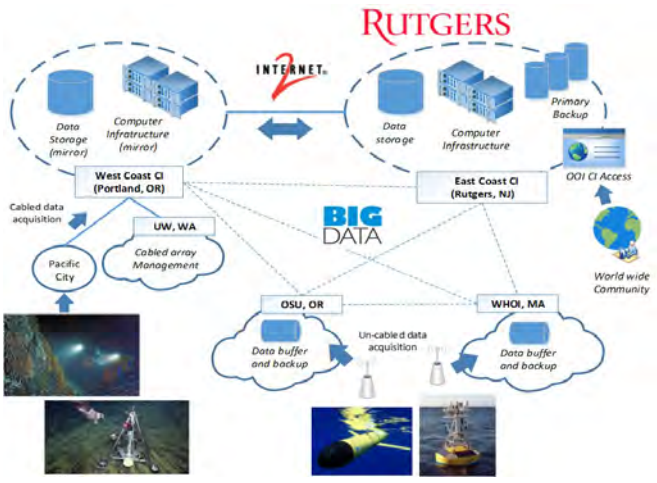


Fig. 1. OOI CI Network Architecture

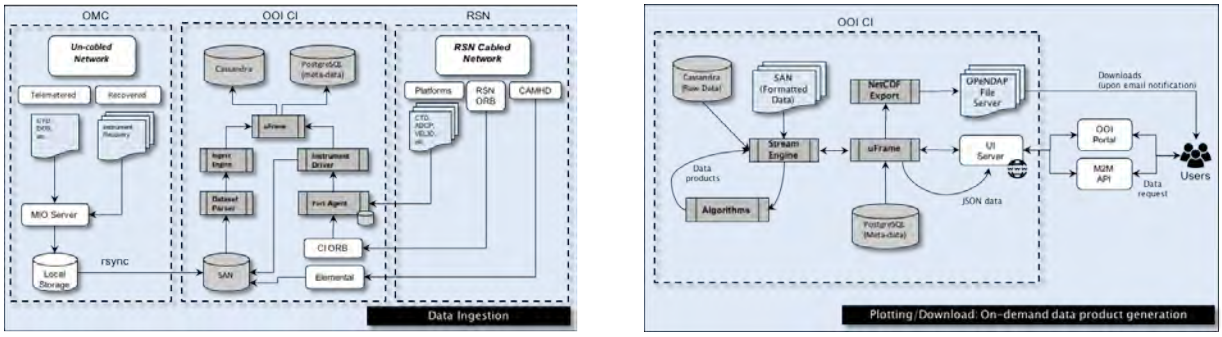


Fig. 2. UFrame-based OOINet software data workflow (left: data ingestion, right: data plotting/download)

(d) Public Data Access: The OOI CI software ecosystem (OOINet) employs the uFrame software framework that processes the raw data and presents it in visually meaningful and comprehensible ways in response to user queries, which is accessible over the Internet through the CI web-based portal access point. A machine-to-machine (M2M) API provides programmatic access to OOINet through a RESTful API. In addition to the portal and API, OOI CI provides the following data delivery methods: (1) *THREDDS Data Server*: delivers data products requested through the CI portal (i.e., generated asynchronously); (2) *Raw Data Archive*: delivers data as they are received directly from the instrument, in instrument-specific format, and (3) *Alfresco Server*: provide cruise data, including shipboard observations. OOI CI software ecosystem permits 24/7 connectivity to bring sustained ocean observing data to a user any time, any place. Anyone with an Internet connection can create an account or use CILogon and access OOI data.

DESING AND IMPLEMENTATION ISSUES

The OOI CI design and implementation principles are based on industry best practises for the different aspects of the CI. The approach is based on a decentralized but coordinated architecture, which is driven by requirements, e.g., data storage capabilities, system load, security, etc.

(a) Redundancy and resiliency: The OOI CI is a mirrored infrastructure for high availability, disaster recovery and business continuity. It implements a resilient information life-cycle management architecture that integrates redundant enterprise storage area network (disk-based) and a robotic library (tape-based). Redundancy is implemented at different layers, for example, an enterprise-level storage network of multiple hard drives managed by an intelligent device manager, reduces the data footprint by reducing data duplication while maintaining data integrity and access performance through storage redundancy, and tape storage, a "last tier" storage that is not dependent on power or cooling, supports longer-term backup and archiving, disaster recovery, and data transport.

(c) Service-oriented Architecture: The core of the OOI CI software ecosystem (Uframe-based OOINet, see Figure 2) is based on a service oriented architecture, a set of data dataset, instrument, platform drivers and data product algorithms, which plug in to the uFrame framework. Uframe-based OOINet uses latest generation technologies for big management data such as Apache Cassandra, which is a state-of-the-art, scalable and highly available distributed database management system designed to handle large amounts of data. Uframe-based OOINet services are exposed through a RESTful API and are available as the M2M interface for external access through a secure endpoint. The use of a well-defined API based on standard protocols enables other systems to interface and interact with OOI CI programmatically.

(c) Cyber-security: The system is based on a multi-tier security approach with dedicated and redundant (highly available) appliances at the CI perimeter. The OOI CI implementation supports encryption of traffic, network traffic segregation, multi-layer traffic filtering, multi-layer access control and comprehensive monitoring. Further, data delivery to external users is implemented through dedicated and distinct storage appliances (i.e., physical and logical isolation from core storage infrastructure) In addition to implementing industry best practices, the OOI CI cyber-security effort includes a comprehensive cyber-security program based on engagement with the NSF Center for Trustworthy Scientific Cyber-Infrastructure. This program encompasses a set of policies and procedures. Regular vulnerability scans/audits (internally and externally) are also performed to the OOI CI.

CONCLUSION

OOI CI has initiated its operational phase and data (including science, engineering and data products) flowing from those instruments is freely available to users. The OOI CI portal provides all data, metadata and data processed via conventional algorithms or direct retrieval from OOI storage or data archives. Data quality and data management will utilize generally accepted protocols, factory calibrations and at sea calibration procedures.

During its early operation (1.5 years), OOI community has been growing every day and is made up of a diverse set of users from 180 different organizations from around the world. At least 500 people has already registered on the OOI Data Portal, which has over 3,000 unique visitors each month[1].

[1] OOI is a NSF-funded effort and involves teams from Consortium for Ocean Leadership, Woods Hole Oceanographic Institution, Oregon State University, University of Washington, Rutgers University, and Raytheon. This document summarizes the contributions from these teams. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

The NSF Cybersecurity Center of Excellence: Large Facilities Services

James A. Marsteller, Chief Information Security Officer, Pittsburgh Supercomputing Center, NSF CCoE Co-PI.
Von Welch, Director, Center for Applied Cybersecurity Research, Indiana University, NSF CCoE PI,

June 19, 2017

Overview of the NSF CCoE

The genesis of the NSF Cybersecurity Center of Excellence (trustedci.org) is with a series of two workshops, the Scientific Software Security Innovation Institute (S3I2) workshops. The S3I2 workshops, held in 2010 [1] and 2011 [2], included representatives of 35 major NSF-funded projects. The original goal of the workshops was to explore a software institute focused on IT security for the NSF community. What the workshops found is that the NSF community faces strong challenges in obtaining access to IT security expertise. Projects are forced to divert their resources to develop that expertise, address risks haphazardly, unknowingly reinvent basic cybersecurity solutions, and struggle with interoperability. The workshops further determined the need for access to expertise was more critical than any new software product.

In 2012, based on these workshop findings, the NSF funded the Center for Trustworthy Scientific Cyberinfrastructure (CTSC) to provide security expertise to the NSF community. Building on the success of CTSC, the NSF Cybersecurity Center of Excellence (CCoE) was funded in 2016 as an expansion of the CTSC. The CCoE draws is a collaboration of four internationally recognized institutions: Indiana University, the University of Illinois, the University of Wisconsin-Madison, and the Pittsburgh Supercomputing Center.

CCoE Services in support of Large Facilities

Science projects manage a number of risks to their scientific missions including risks typically managed by cybersecurity, i.e. malicious entities who attack IT infrastructure to further their own ends at the expense of legitimate users or to explicitly harm those users. To be effective cybersecurity must be tailored for the science community, taking the community's risks, tolerances, and technologies into account. The CCoE's mission is to provide the NSF Large Facility community expertise in cybersecurity for science This mission is accomplished through one-on-one engagements with projects to address their specific challenges; education, outreach, and training to raise the state of security practice across the scientific enterprise; and leadership in advancing the overall state of knowledge on cybersecurity for science through applied research and community building. Examples of these mechanisms follow. Details can be found on trustedci.org.

One-on-one engagements:

- DKIST: DKIST and the CCoE collaborated to develop a cybersecurity planning guide for DKIST that addresses these terms and conditions, aligns with existing institutional

policies, and can be implemented within DKIST's budgetary limitations. This guide was made generally available for other NSF large facilities and projects [3].
- LIGO: The CCoE, LIGO and the Open Science Grid collaborated to establish an international identity federation in support of LIGO's scientific mission.
- Icecube, LSST, NEON: The CCoE helped with the development, assessment, and improvement of operational cybersecurity programs.
- Globus, Pegasus, OSG: The CCoE provided software security consulting and assurance evaluation to helping the NSF community develop more secure software and assess software they are using (or considering using).

Education, outreach and training:

- Situational awareness: The CCoE provides situational awareness of the current cyber threats to the research and education environment, including those that impact scientific instruments, by providing timely email notifications about relevant software vulnerabilities.
- Webinars: The CCoE offers a monthly webinar series to allow NSF projects to share findings and experiences with each other.
- Training: The CCoE regularly provides training, tailored to the science community, on a number on a number of topics, including log analysis, incident response, federated identity management, and developing a cybersecurity program.

Advancing the state of knowledge through applied research and community building:

- Large Facility Security Working Group: to develop a working relationship between those responsible for cybersecurity across the LFs and to advance the development and implementation of best practices, standards and requirements within the community.
- NSF Cybersecurity Summit for Large Facilities and Cyberinfrastructure:The CCoE organizes this annual event to bring together leaders in NSF cyberinfrastructure and cybersecurity to build a trusting, collaborative community, and to address that community's core cybersecurity challenges.

References

[1] William Barnett, Jim Basney, Randy Butler, and Doug Pearson, "Report on the NSF Workshop on Scientific Software Security Innovation Institute (S3I2) (2010)," Oct. 2010 [Online]. Available: https://security.ncsa.illinois.edu/s3i2/s3i2-workshop-final-report.pdf

[2] William Barnett, Jim Basney, Randy Butler, and Doug Pearson, "Report of NSF Workshop Series on Scientific Software Security Innovation Institute (S3I2) (2011)," Oct. 2010 [Online]. Available: https://security.ncsa.illinois.edu/s3i2/S3I2WorkshopReport2011Final.pdf

[3] Jim Marsteller, Craig Jackson, Susan Sons, Jared Allar, Terry Fleury, Patrick Duda, "Guide to Developing Cybersecurity Programs for NSF Science and Engineering Projects, v1," Center for Trustworthy Scientific Cyberinfrastructure, Aug. 2014 [Online]. Available: https://scholarworks.iu.edu/dspace/handle/2022/20026. [Accessed: 18-Jun-2017]

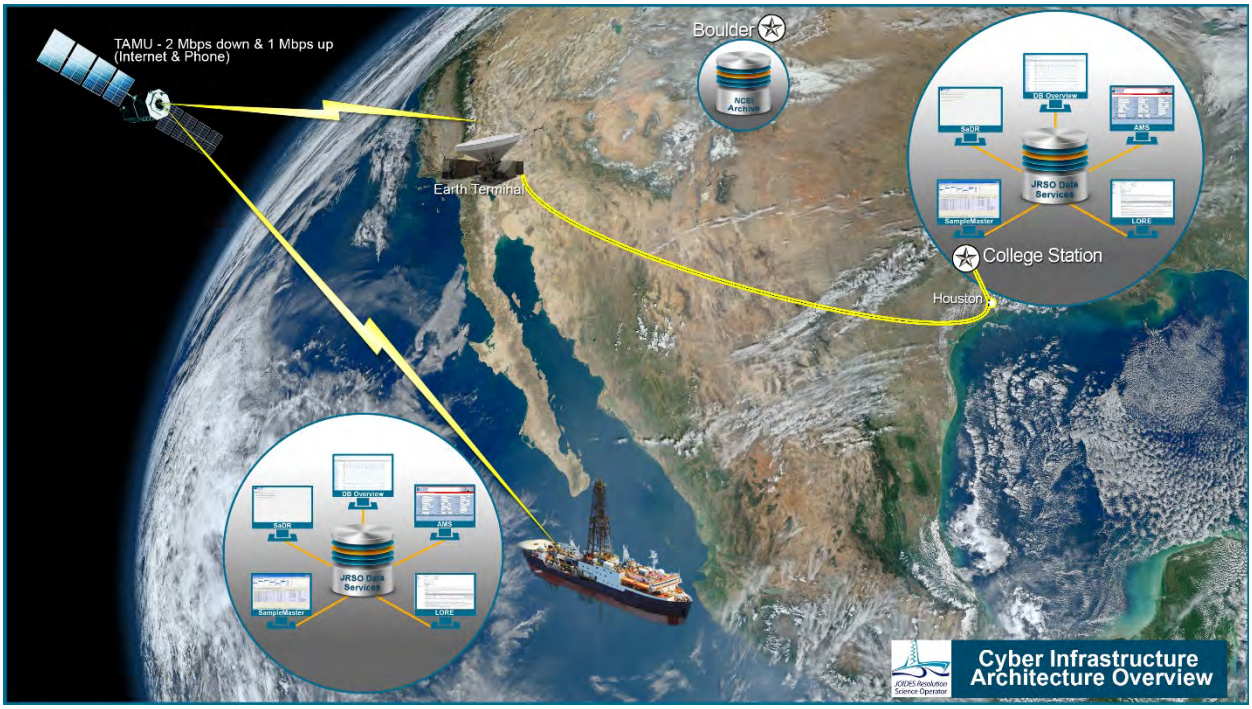
JOIDES Resolution Science Operator Cyberinfrastructure Overview

The *JOIDES Resolution* Science Operator (JRSO) manages and operates the riserless drillship, *JOIDES Resolution*, for the International Ocean Discovery Program (IODP). The JRSO (http://iodp.tamu.edu) is based in the College of Geosciences at Texas A&M University.

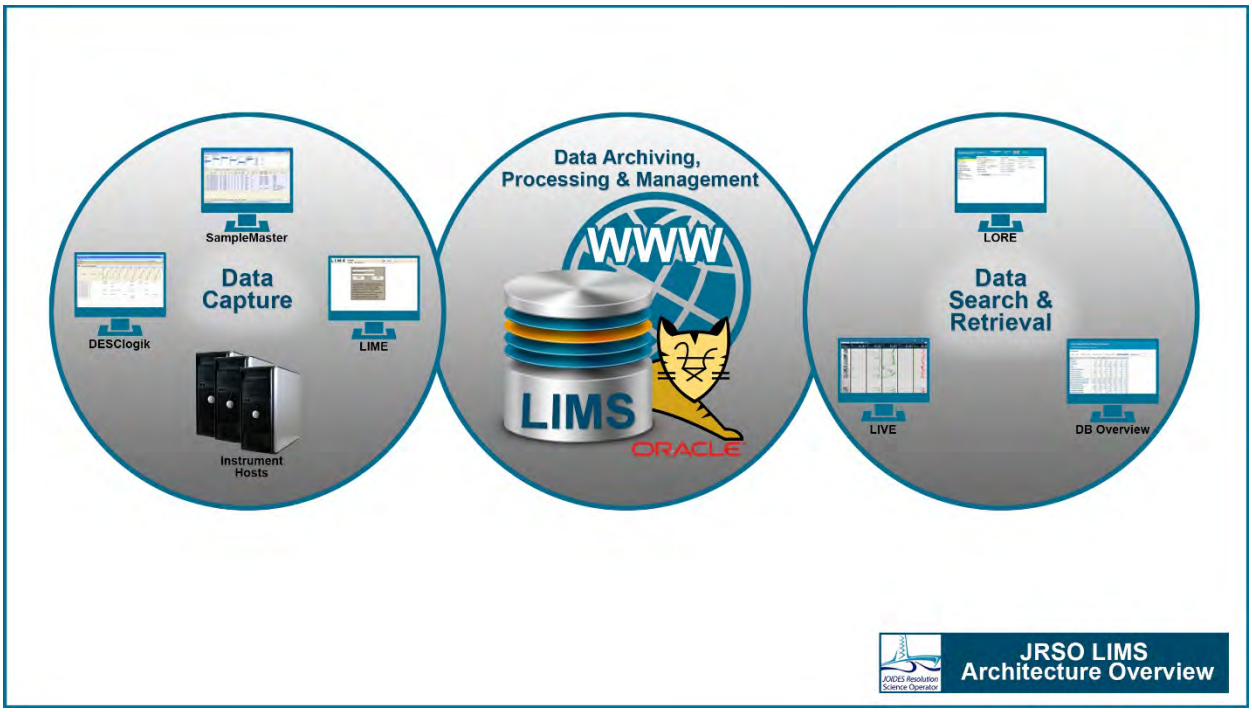
The JRSO is responsible for overseeing the science operations of the riserless drilling vessel *JOIDES Resolution* (JR), archiving the scientific data, samples and logs that are collected, and disseminated via web applications and online publications. The drillship travels throughout the oceans sampling the sediments and rocks beneath the seafloor. The scientific samples and data are used to study Earth's past history, including plate tectonics, ocean currents, climate changes, evolutionary characteristics and extinctions of marine life, and mineral deposits.

The JR is an NSF large facility that serves the global geosciences community. In addition to NSF funding through a cooperative agreement, JRSO operations are partly funded by 22 IODP member nations, including Australia, Austria, Brazil, Canada, China, Denmark, Finland, France, Germany, India, Ireland, Italy, Japan, Korea, Netherlands, New Zealand, Norway, Portugal, Spain, Sweden, Switzerland, and the United Kingdom.

The cyberinfrastructure team supports a split based operations construct, providing cyberinfrastructure, cybersecurity and data management services at sea on board the JR and on shore in College Station, TX. VSAT (very small aperture terminal) satellite services are used to provide connectivity services between ship and shore. Currently, this is a dedicated asynchronous wide area network circuit offering 2 Mbps down to the ship and 1 Mbps up.



The JRSO's Laboratory Information Management System (LIMS) architecture (see picture below) is designed to capture, archive, process, manage, and disseminate data using several JRSO-developed instrument uploaders, client applications and web application tools. LIMS comprises the database that stores the data, the web services that pull and push the data, and the applications and hardware that capture and disseminate the data. One JRSO goal is to make this data, along with the data stored a legacy system (JANUS), more human and machine discoverable. JRSO is hopeful that the NSF-funded Open Core Data project will soon provide the data discovery capability it is seeking.



The cyberinfrastructure team serves approximately 115 internal JRSO staff, 150 international scientists who sail on the JR each year, and the broader global geosciences community.

Under its capital equipment replacement program, the JRSO routinely updates infrastructure services on ship and shore (i.e., servers, storage, backup services, battery backup, and high-speed network). The median age for JRSO infrastructure equipment is approximately six years.

JRSO leverages Texas A&M University policies and tools to maintain its cybersecurity program. JRSO conducts a security self-assessment once per year using RSA Archer GRC in order to remain in compliance with university and state regulations.

JRSO science data is permanently achieved at the NCEI facility in Boulder, CO.

NATIONAL SUPERCONDUCTING CYCLOTRON LABORATORY

The overall mission of the National Superconducting Cyclotron Laboratory (NSCL) at Michigan State University is to provide forefront research opportunities with stable and rare isotope beams. A broad research program is made possible by the large range of accelerated primary and secondary (rare isotope) beams provided by the facility. The major research thrust is to determine the nature and properties of atomic nuclei, especially those near the limits of nuclear stability. Other major activities are related to nuclear properties that influence stellar evolution, explosive phenomena in the cosmos (e.g. supernovae and x-ray bursts), and the synthesis of the heavy elements; and research and development in accelerator and instrumentation physics, including the development of superconducting radiofrequency cavities and design concepts for future accelerators for basic research and societal applications. In all activities an important part of the NSCL program is the training of the next generation of scientists. Upon completion of the DOE-funded Facility for Rare Isotope Beams (FRIB), the laboratory will transition to programs with beams from this facility.

NSCL operates two coupled cyclotrons, which accelerate stable ion beams to energies of up 170 MeV/u. Rare isotope beams are produced by projectile fragmentation and separated in-flight in the A1900 fragment separator. For experiments with high-quality rare isotope beams at an energy of a few MeV/u, the high-energy rare isotope beams are transported to a He gas cell for thermalization, and then sent to the ReA linear post-accelerator for reacceleration. Rare isotope beams in this energy range allow nuclear physics experiments such as low-energy Coulomb excitation and transfer reaction studies as well as for the precise study of astrophysical reactions. The facility has produced over 904 rare isotope beams for experiments, and 65 new isotopes have been discovered at NSCL.

NSCL is a national user facility and has a large user community with over 800 actual, active users in a given year. Most experiments conducted at NSCL involve international collaborations with about 75% of the experiments lead by a US spokesperson.

NSCL provides beams to approximately 30 experiments per year. Experiments are short (~3-7 days) with many changes during and in between experiments. Data acquisition and analysis and simulation framework need to support fast online decision making. Experiments have increased significantly in complexity with an increase of the number of channels read out, often together with high-resolution digitized waveform data. Each experiment can generate up to 10 TB of experimental data set. Storage and backup systems must match such data sizes. Data sets are analyzed on-line during the data acquisition and later off-line either at NSCL or at the spokesperson's institution. Experiments with in-house spokespersons require long-term storage (usually a few years) of the full data set and adequate computing resources for analysis. A computing cluster in the order of 1000 cores dedicated for online analysis is foreseen. Network bandwidths of 100 Gbit/s will be required. External data transfer capabilities must continue to accommodate the needs of a large and distributed user community with increased data set sizes. Data sets are provided to experimenters via magnetic tape, though other methods are available.

NSCL CI supports and enables the Laboratory overall mission. CI includes a broad range of functional areas: business support information technology, networking, accelerator controls, experimental controls and DAQ, and offline simulation and analysis. Internally developed and commercial solutions are used. Systems are primarily managed and maintained by Laboratory

personnel. CI challenges include increasing security requirements, Laboratory growth with FRIB planning and construction, and increasing and foreseen experimental needs.

The Business IT department provides a range of enterprise IT services directly supporting business processes including an internally hosted ERP suite and other customized COTS solutions. Windows based services including Active Directory, Exchange, SharePoint are deployed. More than 500 Windows desktop PCs are maintained.

Business IT department also maintains the Lab-wide network, servers and storage used by DAQ and NSCL Controls and is responsible for overall IT security.

Internet is provided via MSU with MSU assisting with Internet security. Laboratory wired networks are managed internally with MSU supporting wireless access.

The Controls department is responsible for hardware and software controls for accelerators, beamlines, and other experimental equipment. The controls system uses EPICS protocols with graphical monitoring using CS-Studio. NSCL personnel are active in development of both projects. A number of associated systems provide alarms, access controls, archiving etc. for EPICS.

With construction of the FRIB accelerator progressing, new accelerator and cryogenic controls networks are being deployed. These are also EPICS based. The designs emphasis security with FRIB Controls network isolated from other Laboratory systems.

In house developed software forms the core of the DAQ systems. NSCLDAQ is a modular system supporting a range of experiment arrangements. SpecTcl is a compatible analysis software. DDAS is an internally developed digital-DAQ, supporting XIA Pixie-16 Digitizer and compatible with NSCLDAQ. As a user facility, NSCL provides DAQ assistance to visiting experimenters. Typical experiments produce approximately 100 GB of data per day with experiments storing digitized waveforms producing ~1 TB per day. Currently, most experiments' needs are met with 1GE networking and several DAQ computers. Data is recorded to ZFS/Linux servers. Reliability is critical as experiments' beam times are generally limited for less than one week. Visiting experimenters may make use of DAQ systems while present at NSCL.

Increasingly, flexible CPU and software systems are used for DAQ. One purpose is distinguishing overlapping waveform signals from higher rate experiments. The GRETINA experiment is active at NSCL currently utilizing a dedicated farm of approximately 100 PC nodes (1000 cores) for selecting events based on digitized waveforms.

Offline simulations and analysis systems are provided for Laboratory students, faculty and staff. Clustered interactive Linux hosts and a small (~50 node) Linux SLURM batch system are available. Approximately 1 PB of networked research storage is available using ZFS/Linux systems with NFS. Increasing detector complexity, data volumes and analysis complexity require increasing simulation and analysis capacity. Free and widely used applications such as ROOT and GEANT are the norm.

IceCube Computing Infrastructure Overview

Contact: Gonzalo Merino, gonzalo.merino@icecube.wisc.edu
Computing Facilities Manager

IceCube is a neutrino detector built at the South Pole by instrumenting about a cubic kilometer of ice with 5160 light sensors. It uses Cherenkov light, emitted by charged particles moving through the ice to realize the enormous detection volume required for detecting neutrinos. One of the primary goals for IceCube is to elucidate the mechanisms for production of high-energy cosmic rays by detecting high-energy neutrinos from astrophysical sources. The Detector construction started in 2005 and finished in December 2010. Data taking started in 2006 and it is expected to be operated for at least 20 years. The United States National Science Foundation (NSF) supplied funds for the design, construction, and operations of the detector. As the host institution, the University of Wisconsin-Madison, with support from the NSF, has responsibility on the maintenance and operations of the detector. The scientific exploitation is carried out by an international Collaboration of about 300 researchers from 48 institutions in 12 countries (see http://icecube.wisc.edu).

The IceCube data processing is divided in two regimes: online at the South Pole and offline at the UW-Madison main data processing center. Computing equipment is lifecycle replaced on average every ~4 years at the South Pole and ~5 years at UW-Madison. Several collaborating institutions also contribute to the offline computing infrastructure at different levels. Two Tier1 sites provide tape storage services for the long term preservation of the IceCube data products: NERSC in the US and DESY-Zeuthen in Germany. About 20 additional IceCube sites in the US, Canada, Europe and Asia provide computing resources for simulation and analysis.

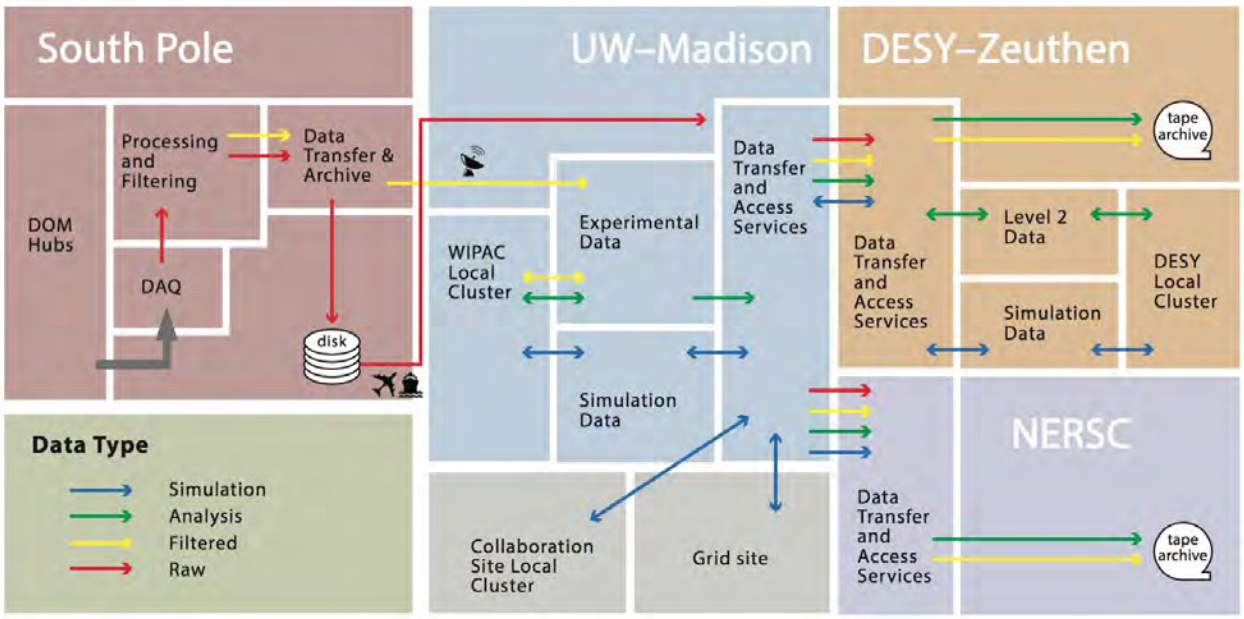


Figure 1 - The IceCube data flow and computing infrastructure.

Online Computing Infrastructure

Aggregation of data from the light sensors begins in the IceCube Laboratory (ICL), a central computing facility located on top of the detector hosting about 100 custom readout DOMHubs and 50 commodity servers. Data is collected from the array at a rate of 150 MB/s. After triggering and event building, the data is split into two independent paths. First, RAW data products are written to disks at a rate of about 1 TB/day, awaiting physical transfer north once per year. In addition, an online compute farm of 22 servers does near-real-time processing, event reconstruction, and filtering. Neutrino candidates and other event signatures of interest are identified within minutes, and notifications are dispatched to other astrophysical observatories worldwide via the Iridium satellite system. Approximately 100 GB/day of filtered events are queued for daily transmission to the main data processing facility at UW–Madison via high-bandwidth satellite links. Once in Madison, filtered data is further processed to a level suitable for scientific analysis.

Offline Computing Infrastructure

The main data processing facility at UW-Madison currently consists of ~7600 CPU cores, ~400 GPUs and ~6 PB of disk. This facility is used mainly for user analysis, but also for data processing and simulation production. Data products that need to be preserved for long time are replicated to two different locations: NERSC and DESY-Zeuthen.

Conversion of event rates into physical fluxes ultimately relies on knowledge of detector characteristics numerically evaluated by running Monte Carlo simulations that model fundamental particle physics, the interaction of particles with matter, transport of optical photons through the ice, and detector response and electronics. Large amounts of simulations of background and signal events must be produced for use by the data analysts. The computationally expensive numerical models necessitate a distributed computing model that can make efficient use of a large number of clusters at many different locations.

Up to 50% of the computing resources used by IceCube simulation and analysis are distributed (i.e. not at UW-Madison). The HTCondor software is used to federate these heterogeneous resources and present users a single consistent interface to all of them:

- Local clusters at IceCube collaborating institutions
- UW campus shared clusters
- Open Science Grid
- XSEDE supercomputers

Overview of UNAVCO

UNAVCO, a non-profit university-governed consortium, facilitates geoscience research and education using geodesy. The website is at http://www.unavco.org.

The UNAVCO consortium membership consists of more than 100 US Full Members and over 80 Associate Members (domestic and international). Through our Geodetic Infrastructure and Geodetic Data Services Programs, UNAVCO operates and supports geodetic networks, geophysical and meteorological instruments, a free and open data archive, software tools for data access and processing, cyberinfrastructure management, technological developments, technical support, and geophysical training. The UNAVCO Education and Community Engagement Program provides educational materials, tools and resources for students, teachers, university faculty and the general public.

Under a 2013 award from the National Science Foundation (NSF), UNAVCO operates the Geodesy Advancing Geosciences and EarthScope (GAGE) Facility. In this role, UNAVCO deploys and operates instrumentation that collects a variety of data to support geodetic with instrumentation systems are deployed globally. UNAVCO provides data management, curation, archiving and distribution services for geodetic data collected or acquired by UNAVCO and by US investigators performing geodesy research with NSF funding. Under certain circumstances non-NSF or NASA funded contributed research data and products are also handled. UNAVCO has been a Regular Member of the ICSU World Data System since 2015.

The Geodetic Data Services (GDS) program manages a complex set of metadata and data flow operations providing a wide range of geodetic/geophysical observations to scientific and educational communities. Sensors currently include Global Navigation Satellite System (GNSS) (downloaded files and high rate data streaming in real time (RTGNSS), borehole geophysics instrumentation (strainmeters, tiltmeters, seismometers, accelerometers, pore pressure and meteorological sensors), long baseline laser strainmeters, and terrestrial laser scanners. Field data are acquired either from continuously operating sites or episodic "campaign" surveys conducted by the community. UNAVCO also acquires and distributes satellite synthetic aperture radar (SAR) data from foreign space agencies. GDS services include data operations (managing metadata; data downloading, ingesting and preprocessing); data products and services (generating processed results and QA/QC and state-ofhealth monitoring); data management and archiving (distribution and curation); cyberinfrastructure; and information technology (systems and web administration). In order to perform this work, GDS maintains a highly specialized technical staff, onsite and offsite computer facilities with networking, servers and storage, and manages a number of sub awards to university groups who provide additional products, software and training.

Key Data and Products

Key data products include GNSS unprocessed and processed data from over 3,000 continuous stations; Terrestrial and Airborne Laser Scanning swaths, point clouds and rasters; raw and processed space borne SAR (Synthetic Aperture Radar) and InSAR (Interferometric Synthetic Aperture Radar) images; borehole strain and seismic data (raw and processed); and raw and processed meteorological observations collocated at selected geodetic stations. Key software developed and supported by UNAVCO for community use include GNSS preprocessing codes, and GNSS data and metadata management software

systems. Through sub awards UNAVCO provides community support for GNSS processing codes.

Facility CI

UNAVCO's CI is intended to provide robust, reliable, secure hardware and software systems that ensure data and metadata integrity from the field sensor to the user. Data are managed through multiple software and systems processes covering acquisition, data communications, ingestion, quality checking, preprocessing and processing, and archiving. Increasingly, web services are used to deliver capability for internal handling as well as discovery tools, visualization, and data delivery processes. UNAVCO maintains internet connectivity with two routes to the outside: a primary link on Internet2 through the Front Range Gigapop, and a failover Comcast commercial Internet link. In-house virtualization with VMWare on newer (less than 5-year old) Dell servers hosts the majority of services; this is supplemented by older Sun server and storage hardware (ten years old); SAN storage technology (Oracle, Infotrend) is supplemented with cloud-based IaaS. A colocation service is used for critical backups and failover capability. The wide range of data types and tools for processing and preprocessing is supported by a variety of software stacks developed starting in the 1990's and evolving through the present with 10 years as the median age. In addition, UNAVCO is investigating deploying several services in the cloud (commercial and NSF XSEDE) through the Earthcube GeoSciCloud project.

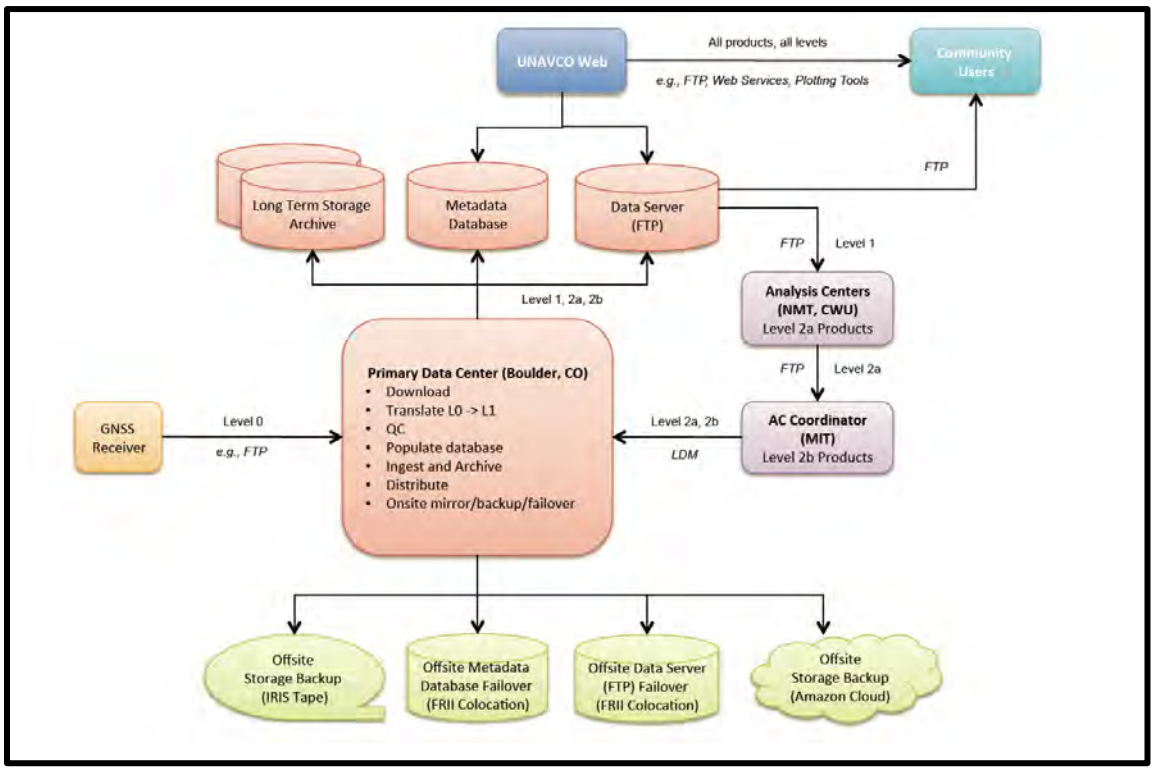


Figure 1. Schematic for UNAVCO's CI for GNSS data and products showing data coming from the field through data and products delivered to users. This schematic, though limited to the GNSS data type, is generally representative of the CI used for other data types (TLS, SAR, borehole strain and seismic, meteorological) handled by UNAVCO.

IRIS Data Services
Tim Ahern
Director of Data Services

The central component of IRIS Data Services (DS) is the IRIS Data Management Center in Seattle, Washington. The DMC relies on other DS components in Albuquerque, La Jolla, University of Washington, LLNL, and Almaty, Kazakhstan to realize its full functionally but the heart of the DS is the DMC. The major CI components are in place at the DMC. We run a fully functional Auxiliary Data Center that is unmanned at LLNL.

1. ***A brief description of the facility, its science mission, and the community (including size, make up – number of individual, number of institutions, etc.). Please include a URL for more information.***

The IRIS DMC is a domain specific facility that meets the needs of the seismological community both within and outside the US. The DMC facilitates science within our domain but does not DO any science. Our science mission can be found in our strategic plan:
http://www.iris.edu/hq/files/programs/data_services/policies/Strategic_Plan_v7.pdf
Our science community numbers in the thousands worldwide.

Mission: *To provide reliable and efficient access to high quality seismological and related geophysical data, generated by IRIS and its domestic and international partners, and to enable all parties interested in using these data to do so in a straightforward and efficient manner.*

IRIS is university consortium with approximately 125 members (US academic institutions with graduate degrees in seismology) and roughly the same number of foreign affiliates scattered all over the globe. We are a 501c3 Delaware corporation. We distribute primary data to roughly 25,000 (3rd level IP address) distinct users or IP addresses per quarter from roughly 12,000 distinct organizations (2nd level IP address). IRIS ingests roughly 75 terabytes of new observable data per year and we project we will more than one petabyte in 2017.

2. ***A description of the key products/services of the facility (data, software, services, etc.)?***

IRIS' primary products are (Level 0, raw and Level 1 quality controlled) time series data. The time series come from roughly 30 types of sensors deployed on/in the ground, in the water column or water bottom, and in the atmosphere. IRIS also produces Level 2 derived products, and manages community developed Level 2 and higher products. (See http://ds.iris.edu/spud/). Level 0 and 1 products are fully documented (metadata) time series data from geophysical sensors distributed globally generated form NSF and other national and international sources. We distribute roughly one petabyte of level 0 and 1 data per year.

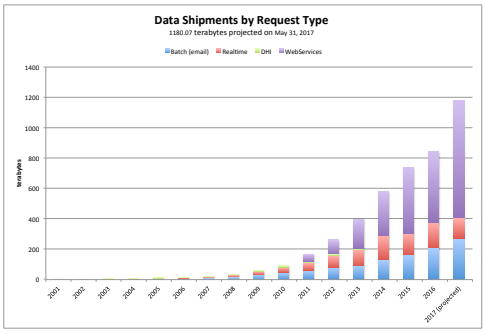


Figure 1 shows volume of time series data shipped from the IRIS DMC to end users and or monitoring agencies since 2001. Major types of shipments include legacy requests in the blue, real time data distribution in the red, and web service distribution in the purple.

IRIS also produces a great deal of community software and offers both IRIS developed and community developed software and tools in Redmine and GitHub repositories. IRIS develops and maintains specific client applications for accessing and working with IRIS data.

All IRIS data assets (Level 0-3) are available through service APIs. Some of the APIs have been adopted internationally (FDSN web services) and other APIs are IRIS developed and maintained and not yet adopted internationally. (see http://service.iris.edu). IRIS also maintains comprehensive documentation and is also the source of documentation for the SEED format, which is the international seismological domain format. (www.fdsn.org)

3. A brief description (including a figure) of the facility CI (e.g., its architecture, key services/components, underlying infrastructure), how it is deployed/distributed, and its operation. What is the median age since deployment of the key CI components.

The figure could be provided at a later date, it is very complex obviously and difficult to provide something at a high enough level as to be useful.

The IRIS DMC operates a primary data center in Seattle as well as an unmanned, fully functional Auxiliary Data Center (ADC) in Livermore California. Major components of CI at the DMC and ADC consist of the following

- **Storage** – IRIS operates large volume Hitachi RAID systems that emphasis storage over performance. We improve performance by indexing the RAID contents in a PostgreSql DBMS. We have roughly 700 terabytes of storage RAID at both the DMC and the ADC. We also operate high performance RAID systems made by NetApp both for reception of real time data and PostgreSql database transactions.
- **Servers**- IRIS runs virtual servers on physica Dell Servers. Virtualization software is VMWare.
  IRS operates Forcepoint Firewalls and A10 Load Balancers. Load Balancers are configured so that a failure at the DMC or the ADC does not remove outsides user's access to services,
- **LANs-** We run 10 gigabit/second LANs sometimes in parallel to form a data backbone internal to the DMC and ADC. We connect to the Internet through the University of Washington.

Storage access to observational data has been abstracted through web services for both internal and external use. Access to data is transitioning from direct SQL access to abstractions thorugh web services. We are very close to running a SOA for both internal and external access.

Our goal is to refresh all major computational and storage hardware infrastructure every four years. Budget pressues sometimes pushes this to 5 years.

We are currently testing operating our software in XSEDE and AWS to see if this is viable.

.

NSF Cyberinfrastructure Facilities Whitepaper

S. Berukoff, E. Cross
Daniel K. Inouye Solar Telescope
National Solar Observatory

Introduction

The Daniel K. Inouye Solar Telescope (DKIST) (http://dkist.nso.edu) is a four-meter, off-axis Gregorian solar telescope currently under construction by the National Solar Observatory and AURA on Haleakala, Maui, Hawai'i. When complete in 2019, it will be the largest solar telescope in the world, providing facility-class, high-resolution solar observations to a small but growing community of students, researchers, and the general public. In full operations, planned to last fifty years, the DKIST will house five complex instruments and a state-of-the-art adaptive optics system, generating over three petabytes of raw data annually. Key to its success, then, is a cyberinfrastructure providing facility and instrument control, scientific and operational data acquisition, and data management, processing, and distribution services. In this whitepaper, we provide a high-level description of primary components of the cyberinfrastructure.

Cyberinfrastructure

The DKIST cyberinfrastructure is comprised of three primary components: the systems and infrastructure providing services to operate the telescope and its supporting subsystems ("Summit"), the core services and infrastructure needed to support science and engineering activities related to observatory operations and network services ("DKIST IT"), and the services and infrastructure performing long-term data management, processing, discovery, and distribution ("Data Center"). These components are highlighted in Figure 1, and discussed in more detail below.

Summit

The DKIST Summit cyberinfrastructure comprises integrated facility, instrument control and safety systems, enabling telescope and dome control, optical alignment and routing, mechanical controls, observation execution and monitoring, instrument data acquisition, management, and distribution, and environmental monitoring and control. These systems are comprised of a High Level Software suite written primarily in Java and Python, utilizing CORBA. They are deployed through configuration-controlled provisioning stacks, including SaltStack, and sit atop an HPC architecture comprising many dedicated nodes interconnected through 10 Gb Ethernet and FDR InfiniBand. The Summit cyberinfrastructure is currently being readied for integration testing as a prelude to observatory integration efforts coming in the next 12-18 months.

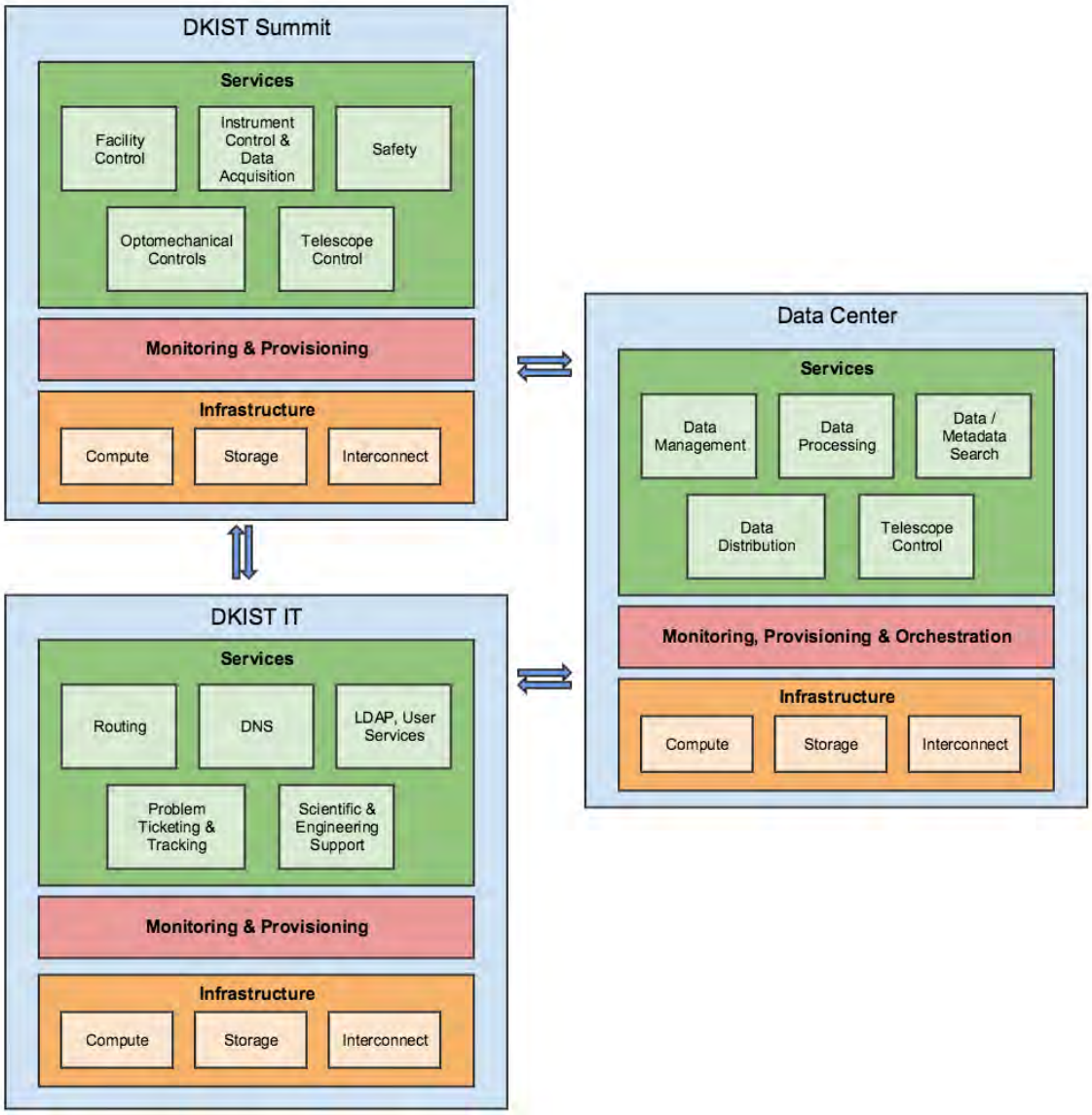


Figure 1: DKIST Cyberinfrastructure

DKIST IT

The DKIST IT supports the observatory through deployment of core services such as routing, DNS, LDAP, and network maintenance and monitoring for the summit and a remote support building, as well ensuring SLAs and/or contracts with partner organizations (U. Hawai'I in Maui and U. Colorado in Boulder at the NSO Headquarters) are met and maintained. In addition, the DKIST IT provides operational support for physical infrastructure (optical fiber, Ethernet and InfiniBand networking, and routing hardware) on the Summit and the remote support building. Services are deployed through configuration-controlled provisioning stacks, sitting atop commodity equipment including Cisco switching. The DKIST IT is ramping its efforts, particularly with regard to network buildout on the Summit and the remote support facility.

Data Center

The DKIST Data Center will provide long-term data management, scientific processing, search, and distribution services for the observatory. It will manage 3.2 PB of data per year, comprised of hundreds of millions of observations and tens of billions of metadata, exported by the Summit and, after calibration, intended for end-user consumption. Thus, data management and processing services must scale effectively with little rework, while data search depends on appropriate data modeling and well-developed use cases to allow end-users to effectively target data of interest. Key aspects of the architecture include a combined microservices and virtual machine deployment, provisioned through SaltStack and managed with Elastic and related tooling. While it is planned for the Data Center to reside at the NSO Headquarters, economies of scale are shifting, indicating a need to ensure "deploy-anywhere" (e.g., commercial cloud providers) can be supported effectively. The Data Center is currently completing its design phase, with development expected to occur in 2018-2020, with phased delivery of critical services occurring as DKIST comes online.

When combined with a rigorous systems-engineering approach, including detailed requirements and interface controls, these three primary components will support DKIST use and scientific data exploitation. Despite the bespoke nature of the Summit CI, there is a significant focus on leveraging open source technologies in the DKIST, rather than relying on integration of commercial products. This is partly due to the long-term nature of the program and tight budgetary constraints. However, there are no free lunches – significant open source adoption without proactive forward replacement planning can leave obsolesced components underpinning critical systems. Given the long development timeline for the DKIST – the first CI work began in 2005 – these issues are already creeping into a yet-to-operate facility. Yet, the state of system development shows significant progress forward, and a bright future, for the DKIST CI.

Summary

This whitepaper briefly discusses the DKIST end-to-end cyberinfrastructure, focusing on the three primary entities and their roles. Each is in a different developmental state, emphasizing the importance of clear requirements and interfaces, effective team communication strategies, and stakeholder management.

Gemini Observatory

White Paper for NSF Cyberinfrastructure Workshop, Sept 2017

Facility Description

The Gemini Observatory consists of twin 8.1-meter diameter optical/infrared telescopes located on two of the best observing sites in the world: Maunakea in Hawaii and Cerro Pachon in Chile. From these two locations, Gemini's telescopes can collectively provide access to the entire sky. Gemini was built and is operated by an international partnership of five countries including the United States, Canada, Brazil, Argentina and Chile. These Participants and the University of Hawaii, which has regular access to Gemini, each maintain a "National Gemini Office" to support their local users. Any astronomer in these countries can apply for time on Gemini, which is allocated in proportion to each Partcipant's financial stake. For the US, Gemini provides the largest publicly-accessible optical/infrared telescopes.

Formally, the Mission Statement is "To advance our knowledge of the Universe by providing the international Gemini Community with forefront access to the entire sky." Gemini's achieves this by supporting peer-reviewed science proposed by the astronomical communities in the participating nations, and providing competitive instrumentation and observing modes in doing so. Over the five-year period between 2012 and 2016, more than 1000 individual Principal Investigators applied for Gemini observing time, from more than 300 academic institutions across the Gemini Partnership.

The Gemini web site: http://www.gemini.edu/

Key products/services

The direct product of Gemini observatory is observational data, taken in appropriate observing conditions, and placed in an archive for access by Principal Investigators (PIs). The service provided to PIs, jointly between the observatory and the NGOs, is to help prepare their observations, then to execute them on the telescopes or support the PI in executing them. Some PIs visit the telescope to make observations, others have their observations taken for them by staff operators. Gemini provides the preparation tool for PIs to create their observations. It also provides a data reduction package for all facility-class instruments. Currently this is based on the standard "IRAF" package distributed by NOAO.

Facility CI

The Gemini Observatory CI (computers, storage and networking; we do not include software in the definition) addresses the combined requirements of telescope operations, data handling and administrative support functions. Each of the four Gemini sites operates identical key services; a redundant core network service to support the distributed network environment, a redundant data storage system capable of replicating data offsite/cross-site in real time, a virtual machine cluster, a physical server farm, a virtual tape library backup environment, which also replicates data offsite, and

instrumentation support infrastructure - such as per-instrument server hardware, network connectivity, remote power management and system monitoring.

The two main Gemini sites (Gemini North and Gemini South) are connected via site-to-site VPN tunnels, that utilize the Internet 2 network infrastructure in the US, with interconnections to the REUNA research network in Chile.

Additionally the two base facility sites in La Serena, Chile and Hilo, Hawaii are equipped with high power computers. These units offer Gemini scientist the possibility of efficiently processing data locally to support their research. While for the most part the consumption of these key services and components is separated, non-operational functions, such as research, project and document management, telecommunications and internet access, enjoy the benefits of increased redundancy and high availability.

The median age of these key CI components is largely dictated by the manufacturers recommendations and enterprise support capabilities and experience in the field. These numbers are in turn transposed to the observatories longevity/obsolescence plan and are therefore understood in advance of the budget cycles. The networking equipment, for example, has a general operating age of around eight years, at which point the support contracts are no longer offered and spares are difficult to procure. The current core network hardware was replaced in 2014 and is set to be replaced in 2022. Similar examples can be made for each key CI component within Gemini, ensuring that the technology will also meet the observatory’s long term requirements.

National Ecological Observatory Network (NEON)
Cyberinfrastructure Overview (July 2017)

Science Mission: Through a Cooperative Agreement with the National Science Foundation, Battelle is constructing the National Ecological Observatory Network (NEON) as a research platform designed to study the biosphere at regional and continental scales and to conduct real-time ecological studies at the scales required to address grand challenges in ecology. www.NEONScience.org

Facility Description: NEON is a new nationwide, "shared-use" research platform of field-deployed instrumented towers and sensor arrays, sentinel measurements, specimen collection protocols, remote sensing capabilities, natural history archives, and facilities for data analysis, modeling, visualization, and forecasting. NEON assets are managed with a cyberinfrastructure of networked processing routines, repositories, and interfaces. The Observatory also supports multi-sensor aircraft payloads (AOPs) operated from leased Twin Otter aircraft, and five mobile deployment platforms (MDPs) that contain both terrestrial and aquatic instrumentation. NEON construction will be completed within the next year.

Key Products & Services: The continental-scale cyberinfrastructure serves 181 data products from 20 regional eco-climatic domains which consist of terrestrial, aquatic, and aerial sampling from over 350 staff. To enable researchers to answer major ecological questions, NEON collects data on a suite of biotic and abiotic variables. As a national research platform, infrastructure, sampling methods, and measurements are being standardized and provided via extensive metadata associated with each downloadable data product. Consistency in collection across locations, through the use of standardized sensors, protocols, and processes, is required to ensure the validity and usability of NEON data by the scientific community and other stakeholders. NEON staff, in concert with automated procedures, evaluate data quality.

The NEON cyberinfrastructure includes models and related computational resources for delivering a range of value-added "data products" based on the in-situ, experimental, and remote sensing components. These models and algorithms perform quality control processing, classification, scaling and interpolation functions, as well as provide a platform for external researchers accessing the data to detect patterns, test hypotheses, and project ecological forecasts against seamless, continental scale data layers.

The cyberinfrastructure, which is headquartered in Colorado, publishes both real-time provisional data, and annual releases of observatory-wide versions of results. The cyberinfrastructure architecture is built across facilities which range from the central, commercial data center, to headquarters development environments, to cloud-based data acquisition/staging applications, to distributed sites with dedicated local unmanned facilities, communications, routing controls, and local data logging. Repository content is managed via a central object store, a portfolio of relational databases, and shared code libraries. The cyberinfrastructure includes numerous operational subsystem including: ingest; archival; calibration; processing pipelines; metadata management; specimen custody management, and publishing functions. NEON's web presence consists of interactive portals to data assets, community services, and application programming interfaces (API).

The cyberinfrastructure development team uses best practices approaches to software development via an iterative approach to development (using industry-standard Agile methodology) that stresses the evolving nature of requirements gathering and development. The team emphasizes best practices engineering principles, including code re-use and definition of interfaces to facilitate object-oriented software integration and provide a basis for future growth. Formalized QA methods are applied to in unit, integrated, and regression testing. Segregated development, test, integration, and production environments control releases. The NEON cyberinfrastructure is designed to invite incremental improvements through incorporation and testing of open-source code from community members.

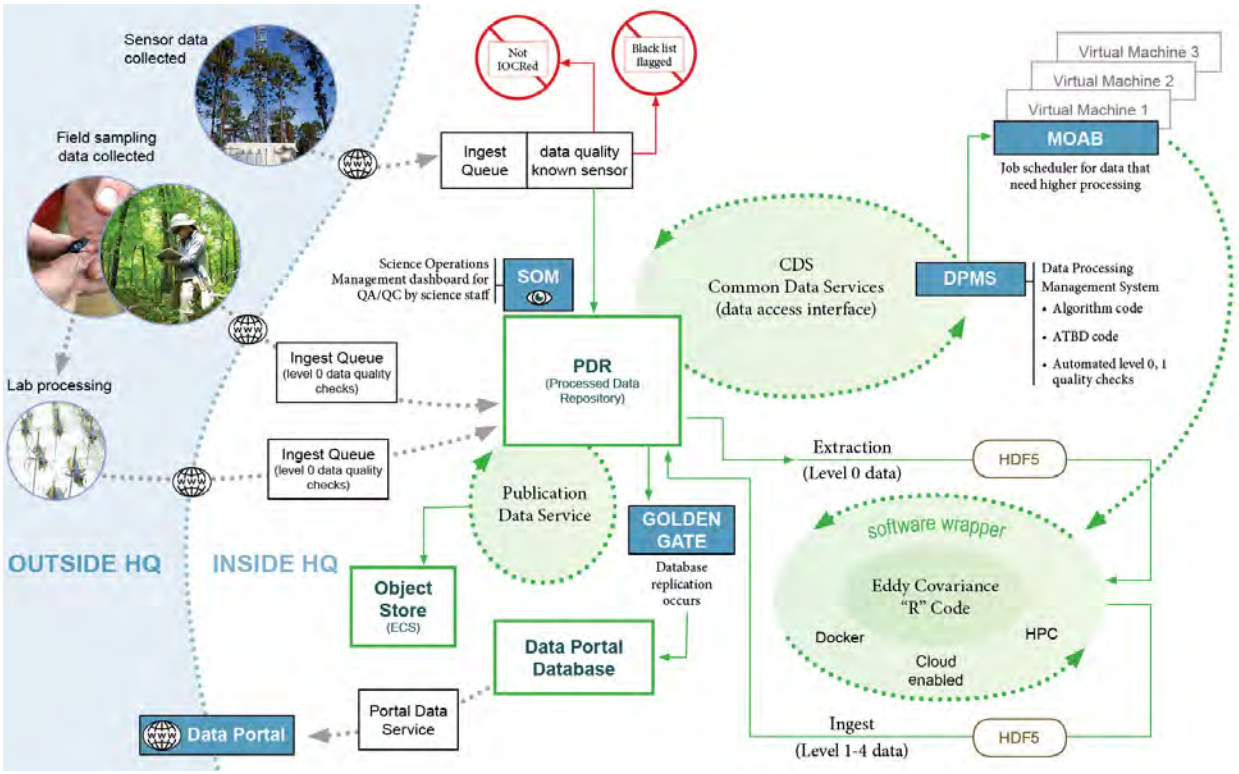


Figure 1. Generalized landscape of data flow through the NEON cyberinfrastructure.

Management & Community Engagement: Leadership is conducted from the NEON Project headquarters in Boulder, Colorado, where core science, management, and administrative functions for the Observatory is managed through the 30-year operational life. NEON's operation is periodically adapted through guidance from the Science, Technology, and Education Advisory Committee (STEAC). Community input is facilitated by 20+ Technical Working Groups. Some NEON products are hosted by community partner organizations: BOLD; SRA; MG-RAST; PhenoCam; AeroNet; AmeriFlux, and DataOne. NEON participants include dozens of laboratories, universities, and agencies. Initial user statistics reflect over 10,000 users from domestic and international organizations.

OCEAN NETWORKS CANADA'S OCEANS 2.0 DIGITAL INFRASTRUCTURE

BENOÎT PIRENNE, DIRECTOR, USER ENGAGEMENT, OCEAN NETWORKS CANADA, VICTORIA, BC

Ocean Networks Canada (ONC) is a world-leading organization supporting ocean discovery and technological innovation. ONC is a not-for-profit society that operates and manages innovative cabled observatories on behalf of the University of Victoria, in British Columbia. These observatories supply continuous power and Internet connectivity to various scientific instruments located in coastal, deep-ocean, and Arctic environments. ONC's arrays host hundreds of sensors distributed in, on and above the seabed along with mobile and land-based assets strategically located. The instruments address key scientific and policy issues (subsea earthquakes and tsunamis, ocean acidification, marine biodiversity, etc.) within a wide range of environments. (See Fig. 1).

Fig. 1: Ocean Networks Canada ocean and coastal observing facilities include the VENUS and NEPTUNE systems off British Columbia (see inset), together with a number of community observatories across Canada (including in the Arctic). With over 400 instruments reporting real-time data from the deep ocean to the coast, ONC's Oceans 2.0 data management system makes big data and products available to scientists, governments, municipalities and first nations.

ONC has built Oceans 2.0, the digital infrastructure that manages vast amounts of complex data streams. Oceans 2.0 is unique in that it supports the continuously increasing volume (currently at 500 terabytes), the variety of data types (dozens of instrument types and over 5000 individual sensors), the data structures that enable rapid access and delivery of analytically-derived alerts, the consistency of data through an instrument management system with robust and rich metadata, as well as automatic and manual QA/QC.

Ocean Networks Canada's Oceans 2.0 sensor network data management can also host and distribute data for 3rd parties, and has features for attribution and access restrictions. Some of its unique data access features include a distributed, live video annotation (SeaScribe) and a video search capability (SeaTube); tools for viewing and searching a hydrophone data archive; tools for the continuous browsing of complex time series data, etc. It also includes an integrated suite of observatory management tools to monitoring and control the infrastructure (electrical, communication and data flow control — see Fig. 2).

Fig. 2: ONC's Oceans 2.0 system offers integrated monitoring and control tools for managing power, data flow and communication with the "Device Console" tool. The navigation follows the tree structure of the network topology. The same interface allows operators control over any part of the infrastructure regardless of where it is located.

Oceans 2.0 is solidly founded on a Service Oriented Architecture based on a core Enterprise

Service Bus. This provides a high performance platform based on a modular, loosely-coupled component architecture, and allows for the simplified addition of the constituent modules on an as needed basis.

With this architectural foundation, Oceans 2.0 provides a simplified, well-defined, event-driven and "pluggable" system which can be scaled as the organization's requirements change. (See Fig. 3).

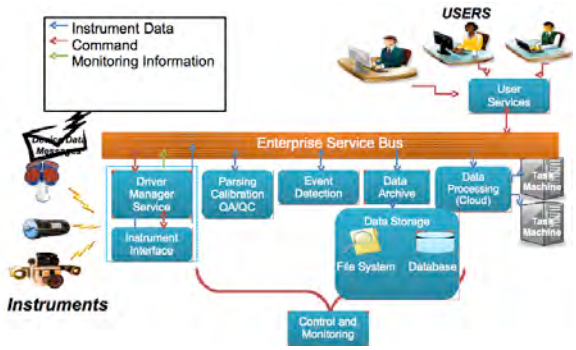


Figure 3. Elements of Oceans 2.0 and their relationship from sensor data generation to the archive and users.

The Oceans 2.0 components include: The *Enterprise Service Bus,* which is the message passing system that allows all parts of Oceans 2.0 to interact and pass information and data. All functional components of Oceans 2.0 use it to asynchronously intercommunicate.

The *Driver Manager Service and Instrument Interface* represent the part of the software that interacts with instruments and their integrated sensors. The software standardizes access to instruments and generalizes their data structures so that they can be used downstream by other software components. Another critically important role of the drivers is their time stamping function that guarantees the same time reference across all the instruments connected to all of the supported networks. Once a raw data record is obtained from an instrument, the driver publishes it to the service bus that subsequently makes it available for other software elements in the system. Oceans 2.0 has drivers for more than 100 different types of instruments from a variety of manufacturers.

Parsing & Calibration, QA/QC is the software module that takes the raw readings from instruments and turns them into meaningful, corrected values, possibly after an optional calibration stage. Moreover, a level 0 automated calibration can be configured to flag sensor values that are out of range.

Event Detection is used to create custom reactions for real-time events. Users can create event definitions using algebraic formulas or other triggers, and associate appropriate reactions if the event occurs. Event Detection currently has several use cases within Oceans 2.0: it is used to perform Quality Assurance and Quality Control (QA/QC) evaluations, and to synchronize acoustic device sampling so as to prevent interference. Another, significantly more advanced event detection system is the ability to detect P-wave from accelerometers, helping with the detection and characterization of earthquakes.

Data Archive takes all data traffic between the instruments and the "surface" side and archives them.

Data Processing indicates the part of the system where data products are generated from the raw data. These include data format conversion, plots and images, etc.

User Services includes a combination of data access and visualization tools, using either a web interactive interface, an application programming interface consisting of standard-abiding web services and a "sandbox" where users upload data processing codes and run them.

Security and resilience. The security of the system against malevolent or accidental access by unexpected parties is provided by isolation of all the key component in secure, private and non-routable networks. The Oceans 2.0 architecture has also been designed around resilience, in particular for the data acquisition component including: fault tolerance in case of network path breakdown, multiple safeguards to minimize data loss in case of unexpected anomalies; and, support of multiple archive centres containing integral data copies.

Americas Lightpaths (AmLight) supporting NSF Large Facilities and CI in the Americas

White Paper for NSF Large Facilities Cyberinfrastructure Workshop

Florida International University (FIU) is the awardee of the NSF International Research Network Connections (IRNC) program, under cooperative agreements, to build and operate the network infrastructure that links the U.S. research networks with peer networks in South America and the Caribbean. This network infrastructure, referred to as AmLight, consists of multiple 10/100 Gbps links, presently totaling 240Gbps of aggregate bandwidth capacity between the U.S. and South America; an international exchange point facility in Miami, Florida, called AMPATH, which terminates the many network connections that depart from the U.S. to, and that arrive from, the research and education networks of the nations of South America, and the Caribbean. FIU has been performing this role on behalf of the NSF since 2005.

Science data flows between NSF Large Facilities or CI, operating in South America or the Caribbean, benefit from the use of the AmLight network links and the network infrastructure that connect these large facilities or CI back to the U.S. AmLight network links are built and operated by network operators whose purpose it is to support research and education communities. The commitment to collaborate and coordinate among the network operators is underpinned by agreements (MOUs) FIU established with the network operators participating in AmLight. For example, in the U.S., network operators are primarily FIU, Florida LambdaRail (regional network in Florida), Internet2 (U.S. national research and education network), ESnet (U.S. national research and education network), and a few others. In South America, network operators are primarily RedCLARA (regional network of Latin America), RNP (national research and education network of Brazil), ANSP (Academic Network of Sao Paulo), REUNA (national research and education network of Chile), and others.

Remote users of NSF Large Facilities in South America or the Caribbean depend on reliable network services to access CI for their research. For example, this could be a low latency network service to remotely control a telescope in Chile, or a higher throughput network service to transfer a large LHC data set from a data center in Sao Paulo to Fermi Lab. Impacts to network services, caused by fiber cuts, power outages, retransmits, etc., will significantly impact applications using CI at NSF Large Facilities. The impact could render the science application inoperative when the NSF Large Facility and the CI are continents apart. For example, a fiber cut will impact a science data flow from an observatory in Chile to the NCSA data center in Champaign, Illinois. Fortunately, networks participating in AmLight have instrumented their networks with monitoring and measurement instruments to detect network impacting events. Data collected from these instruments enable network operators to represent the conditions on the networks that constitute the end-to-end path of the data flow. To inform users of CI at NSF Large Facilities, a web-based interface is available that shows network conditions for many of the interconnection points along the networks between the U.S., South America and the

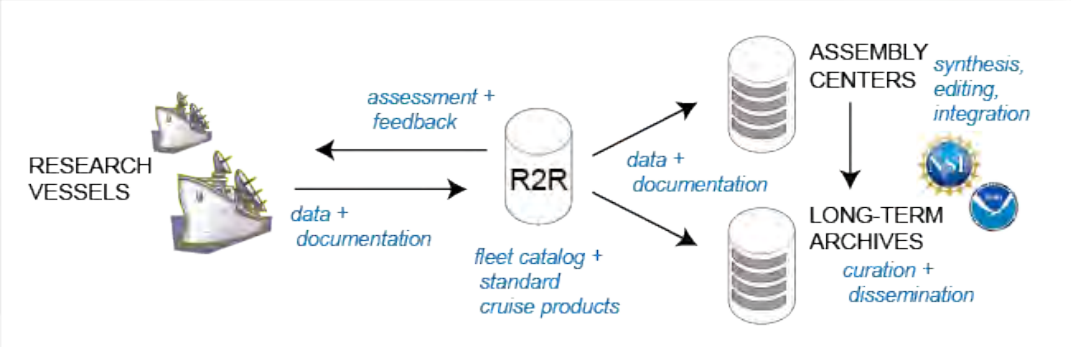
Caribbean. With the web-based interface and other deployed tools, AmLight is achieving its goal to improve detection of network impacting events and to minimize their impacts on science data flows.

Flows of science data between endpoints is a very important unit of measure for AmLight. Flows should experience little to no friction along the end-to-end path. The end-to-end path should be instrumented to monitor and measure network conditions that could impact science data flows. Mechanisms, such as a Science DMZ or Data Transfer Nodes (DTN), should be considered as best practices to reduce friction on science data flows. AmLight can facilitate the implementation and use of these mechanisms for NSF Large Facilities and CI.

The Rolling Deck to Repository Program Marine Data Services for the US Academic Research Fleet

Mission

The Rolling Deck to Repository (R2R; www.rvdata.us) program documents and preserves environmental sensor data acquired during scientific expeditions on U.S. academic research vessels. The program is a collaborative effort between Lamont-Doherty Earth Observatory, Florida State University, Scripps Institution of Oceanography, and Woods Hole Oceanographic Institution; and works closely with the University-National Oceanographic Laboratory System (UNOLS; www.unols.org), whose membership includes 58 U.S. academic institutions supporting oceanographic research worldwide. R2R is funded primarily by the National Science Foundation with support from the Office of Naval Research and the Schmidt Ocean Institute.



Products and Services

Each research vessel delivers a package of original/unprocessed navigational, geophysical, oceanographic, and meteorological data from its permanently installed sensor systems to R2R at the end of an expedition, along with a manifest that provides the vessel ID, cruise ID, title, start/end dates and ports, and science party members. R2R deposits a copy of each package to private offline storage segments at both the NOAA National Centers for Environmental Information (NCEI) and Amazon Glacier. Each package is then broken out into individual datasets according to sensor type, make, model, and file format; and a master catalog of expeditions and datasets is published online via the R2R Web site. A Digital Object Identifier (DOI) is published for each expedition as well as for each dataset. After permission is granted by the Chief Scientist, individual datasets are posted for public download via the R2R Web site. Selected datasets are submitted to NCEI for dissemination and inclusion in global syntheses.

R2R assesses the quality of selected data types, using a scripted workflow and criteria developed in collaboration with specialists in the science community. The assessment results are published online via the R2R Web site, and standard ratings are calculated as part of feedback to vessel technicians. R2R also produces a standard set of data products after each expedition including quality-controlled shiptrack navigation, underway geophysical profiles (gravity, magnetics, bathymetry), water column depth profiles from CTD hydrocasts, and real-time

meteorology/near-surface oceanography; all of which are posted for public download via the R2R Web site. R2R supports an Event Logger application, including a shipboard microserver, to assist science parties in documenting their scientific sampling while underway.

The R2R program interoperates with 14 other data repositories, primarily NSF-sponsored, that manage other kinds of marine data content related to cruises inventoried in the R2R Catalog. The data content hosted in these repositories includes data acquired with specialized science party instruments and national instrument facilities, scientific sampling logs and associated laboratory analyses, as well as processed data products derived from field data, global synthesis products, and links to articles in scientific journals. A suite of Web-based services support interoperability including a OGC Web Feature Service (WFS) that provides shiptrack geometries; a Catalog Service for Web (CSW) that provides ISO 19139 XML records for expeditions; a W3C "Linked Data" graph and associated RDF Query Language (SPARQL) endpoint for Semantic Web clients; and customized Atom+GeoRSS feeds for partner programs such as OOI and ECS. The complete inventory of expeditions and datasets in the R2R Catalog are discoverable in global research indexes such as DataCite (http://search.datacite.org).

Infrastructure

R2R's computer infrastructure is primarily located on the LDEO campus of Columbia University in Palisades, New York, with selective extensions to commercial providers. The LDEO campus cluster consists of six Dell Linux-based servers, six ACNC fiber storage arrays, and a supporting local network of switches/routers, firewalls, and environmental monitors, split between two buildings. Hardware is typically refreshed on a 5-year cycle. Monitoring and backups are implemented via the Nagios and Bacula open-source packages. The application infrastructure is open-source software consisting of Apache Httpd/Tomcat, PostgreSQL, and PostGIS backends. Programming is primarily PHP and Shell scripting, managed in GitHub private repositories, using open-source libraries such as GDAL and MB-System. Commercial provisioning is used for outward-facing Web services such as the R2R Search page (Linode.com), and an off-site backup copies of R2R data packages received from all research vessels are stored in a Amazon Web Services Glacier vault in the US-West-2 (Oregon) zone.

NSF Large Facilities Cyberinfrastructure Workshop

Oregon State University College of Earth, Ocean, and Atmospheric Sciences:

Regional Class Research Vessel Project, http://ceoas.oregonstate.edu/ships/rcrv/

Science Mission: The coastal ocean encompasses the most complex range of oceanic phenomena on the globe. Coastal regions are sensitive to human alteration from water and air pollution, resource extraction, transportation, and recreational activities. Wind- and freshwater-driven coastal ocean flows directly affect regional climate. As conveyors for heat and salt and regions of strong vertical mixing, boundary currents play an outsized role in the large-scale ocean circulation. Vigorous interactions between the coastal ocean and the atmosphere control many biogeochemical processes (e.g., the exchange of macronutrients and micronutrients between the land, ocean, and continental margin sediments).

The coastal oceans are extremely productive, accounting for a large percentage of the world's wild seafood and most of the aquaculture. They are the dominant sites for burial of organic matter, important in net marine uptake of atmospheric CO_2, and locations of major hydrocarbon resources, including oil, gas, and methane gas hydrate. The coastal oceans can be sites of wind and wave energy extraction, play host to the deposition of river sediments, including dredge spoils, and are sites of tectonic activity, including hazardous earthquakes and tsunamis. To better understand such coastal phenomena and their importance in the Earth system, ocean scientists and educators must accelerate exploration and sustained regional observations of marine physical, chemical, biological, and geological processes.

Even with the development of new platforms to study the ocean—such as cabled observatories and underwater robots— coverage is scant, and ships are more vital than ever for multi-disciplinary observations and sampling of the ocean. The RCRVs will feature advanced sensors and sampling systems, and through telepresence capabilities and satellite communications, will bring science at sea to classrooms, the public, and researchers ashore. Oregon State is proud to be leading the charge in developing next-generation vessels that promise state-of-the-art platforms for the nation's scientists and students to explore our ocean planet

Facility size & composition during construction: Core team - OSU = 15 (members), Engineering & Design Support - The Glosten Associates = 4, Science Oversight Committee = 11

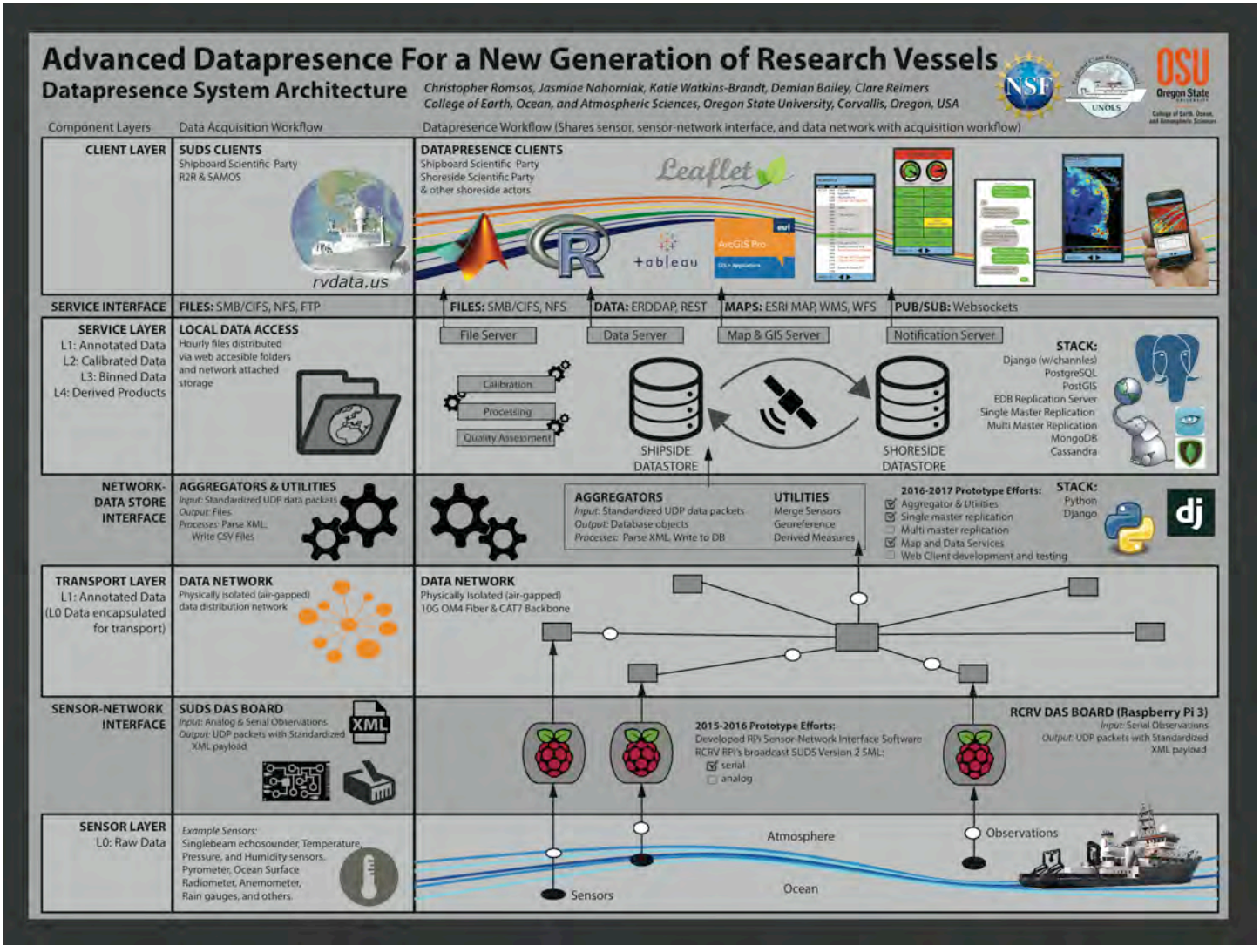
Facility size & composition when operational: OSU Class Management Office = 4, OSU Ship Operations = 15, Institution Two = approx. 15, Institution Three = approx. 15.

Key Products and Services: Oceanographic research ships, and the Regional Class Research Vessels specifically, are the primary platform from which ocean science is conducted. A research vessel must function as observatory, lab, and accommodation. Therefore the facility must provision cyberinfrastructure for the research enterprise, for vessel operations and for quality of life. In addition to these core services the RCRV facility shall also provide a system

for real-time bi-directional transfer of data and information between shipboard and shoreside parties.

To support these requirements we've developed a system for sensor data transmission, capture, archive, replication, and use. The system incorporates a variety of open-source and commercial products including, Enterprise DB Postgres Advanced Server, Apache, Django, Highcharts, Tableau, Mapserver, Leaflet, and ERDDAP.

Facility Cyberinfrastructure: The figure below describes the basic architecture, components, and services of the RCRV Datapresence System. The system is currently under development and testing and has been deployed successfully during prototype cruises.



Acknowledgements

We would like to acknowledge the contributions of Forough Ghahramani, Caroline McHugh and Laura Readie from the Rutgers Discovery Informatics Institute (RDI^2) for their help with the workshop organization. Forough and Caroline also helped with the execution and analysis of the surveys. We would like acknowledge the contributions of Greg Jones from University of Utah and Rafael Ferreira Da Silva from University of Southern California for taking notes at the meeting, and Christine Pickett, and Christine Pickett from University of Utah for editorial help with the report. Greg Jones also contributed significantly to the content of the report. Finally, we would like thank Nathan Galli for putting together the workshop website. The workshop was support by the National Science Foundation through grant number ACI 1742969.