RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY

NSF Large Facilities Cyberinfrastructure Workshop

Ivan Rodero
Rutgers Discovery Informatics Institute (RDI^2)

NSF Large Facilities Workshop
Alexandria, VA, USA, May 1, 2018

ACI 1742969

Workshop Organization

- Steering Committee (Leading experts from the facilities and CI community)
  - Manish Parashar (PI and Chair), Rutgers University and OOI
  - Stuart Anderson, California Institute of Technology and LIGO
  - Ewa Deelman, Information Sciences Institute (ISI), Univ. of Southern California
  - 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

- Acknowledgements
  - **The workshop was support by the National Science Foundation through grant number ACI 1742969**
  - Forough Ghahramani, Caroline McHugh, Laura Readie, Daniel Martin (RDI^2), Greg Jones, Christine Pickett (Univ. of Utah), Rafael Ferreira Da Silva (USC), Nathan Galli

Motivations

- Cyberinfrastructure (CI) is a **critical component** of NSF facilities, and is growing in **scale and complexity**
- Data delivery mechanisms make it harder to **integrate data across multiple facilities** as part of a scientific workflow, resulting in data silos.
- Facilities and CI communities must collectively explore how to **provide** and **sustain** essential CI components and services to meet current and future needs

Workshop Goals

- 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.
- ***Generate recommendations*** that can serve as inputs to current and future NSF CI-related programs.

Workshop Activities

Pre-
Workshop
Post-

- **White papers**
- **Questionnaire**
- **Workshop Website**

[Spring/Summer 2017]

- Panels and Breakouts
- Key Findings
- Recommended Actions

[September 6/7, 2017]

- Post-Workshop Survey
- Workshop Report
- Study of White Papers

[Fall 2017, Spring 2018]

Pre-Workshop Survey

- Whitepapers (up to 2 pages in length)
  - A brief description of the facility, its science mission, and the community
  - A description of the key products/services of the facility
  - A brief description of the facility CI
  - **22 submissions received**
- Questionnaire (8 questions)
- All responses and whitepapers available at ***facilitiesci.org***

Pre-Workshop Survey - Questionnaire

1. Significant components of the CI developed in-house

- Most of the CI components are developed in-house (~60%)
  - Tailored solutions to deal with a particular environment and facility needs
  - Often to deal with data management: sensor data capture, data distribution and replication
  - Monitoring solutions are often customized as well

Pre-Workshop Survey - Questionnaire

2. External CI capabilities and services and/or externally developed tools used

- Reuse of basic software systems
- Use of NSF-funded CI software (e.g., Globus GridFTP) and CI Platforms (e.g., Open Science Grid, XSEDE)
- Some projects are leveraging capabilities delivered by CTSC
- Some facilities are leveraging cloud technologies for data management

Identification of the tools and criteria to select them

- IT Staff, governing committees, technical teams, small groups
- Community involvement
- Use of requirement gathering, evaluation of various existing software solutions, etc.

Pre-Workshop Survey - Questionnaire

3. Most used and most challenging CI components

(1) Data (rapidly growing)
(2) Networking (reliable, high bandwidth, international scale)
(3) Computing (large and diverse workflows, web services e.g., Jupyter)

Aspects shared as best practices

- Use Systems Engineering to manage CI lifecycle and interfaces
- Bake in redundancy to provide high availability

Pre-Workshop Survey - Questionnaire

4. Aspects of the facility CI and its operation seen as challenges or gaps

- Budgets (growing user base with shrinking budgets, maintain CI)
- Recruiting and retention
- Technology / Operations (evolving requirements, migration to Cloud)
- Security

"CI lessons learnt"?

- – Implementation of industry "best practices" for deployment/operation
- – Ability to trace CI features to requirements and business needs
- – Models for communication/interaction

Pre-Workshop Survey - Questionnaire

5. Key risks in facility CI

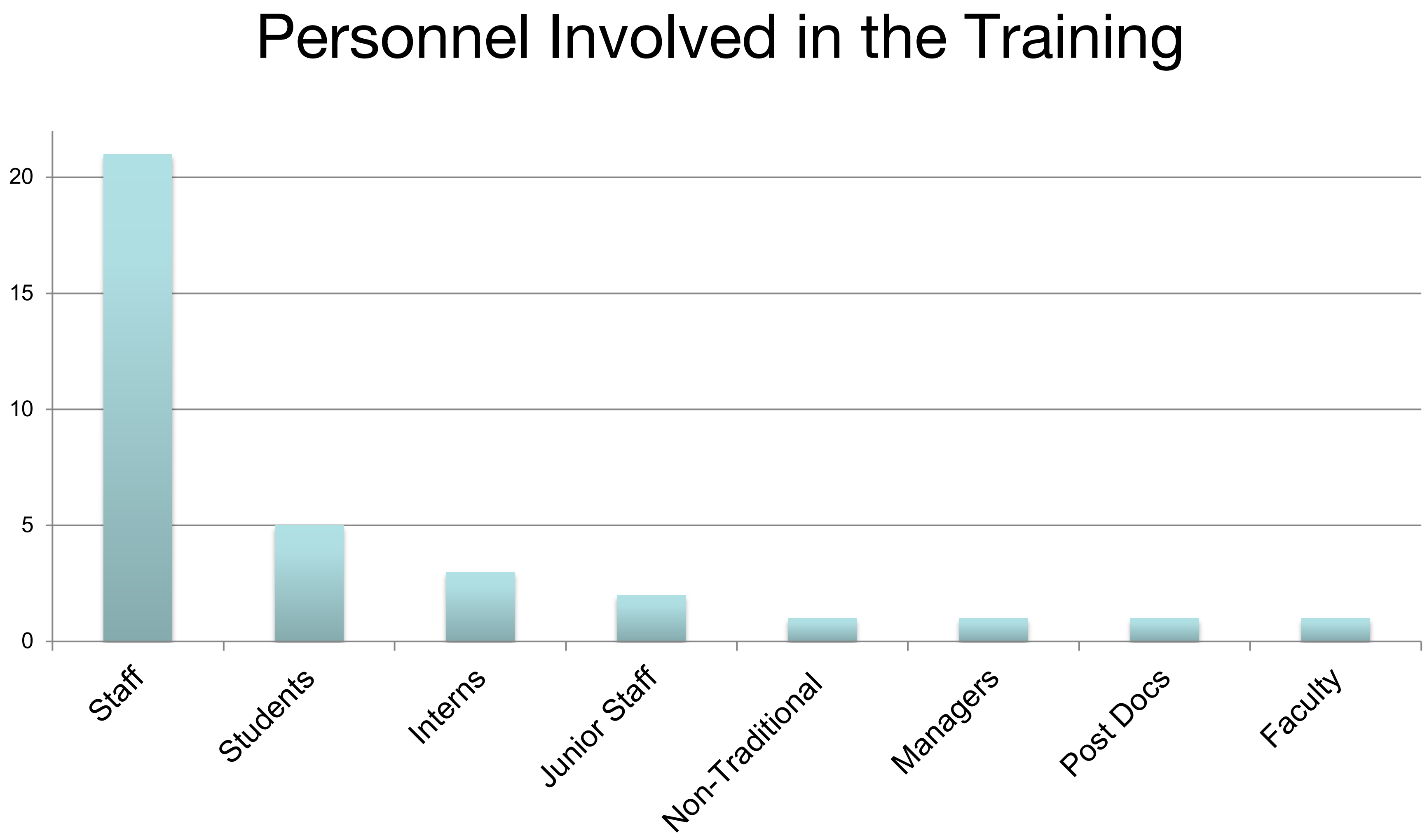
- Funding
- Infrastructure and technology (technology disruptions, sensor vulnerability)
- Workforce (loss of personnel, workforce recruiting and retention)
- Integration / interoperability (sharing knowledge, expertise, infrastructure)
- Scalability (Growing scale and diversity of user community)
- Security (Growing cybersecurity threat landscape)

Pre-Workshop Survey - Questionnaire

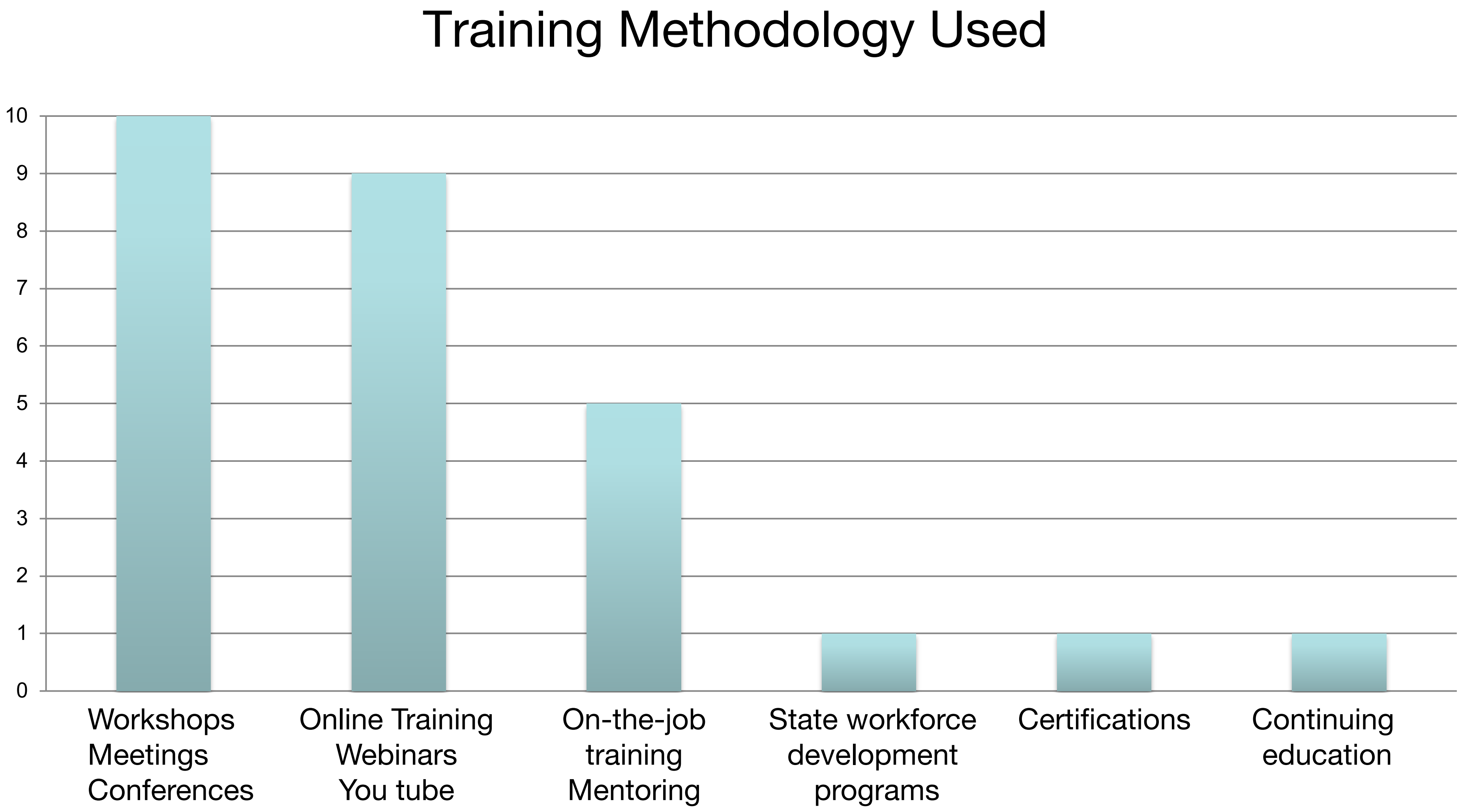
6. CI-related workforce development activities

- Workforce development and retention is considered one of the top priorities and risks
- High variability of approaches ("allow" personnel getting involved in training, attending workshops, etc.)
- Workforce development involves diversity challenges (e.g., variety of seniority/expertise)
- Not clear if there is a systematic budget allocation
- The following slides report some statistics …

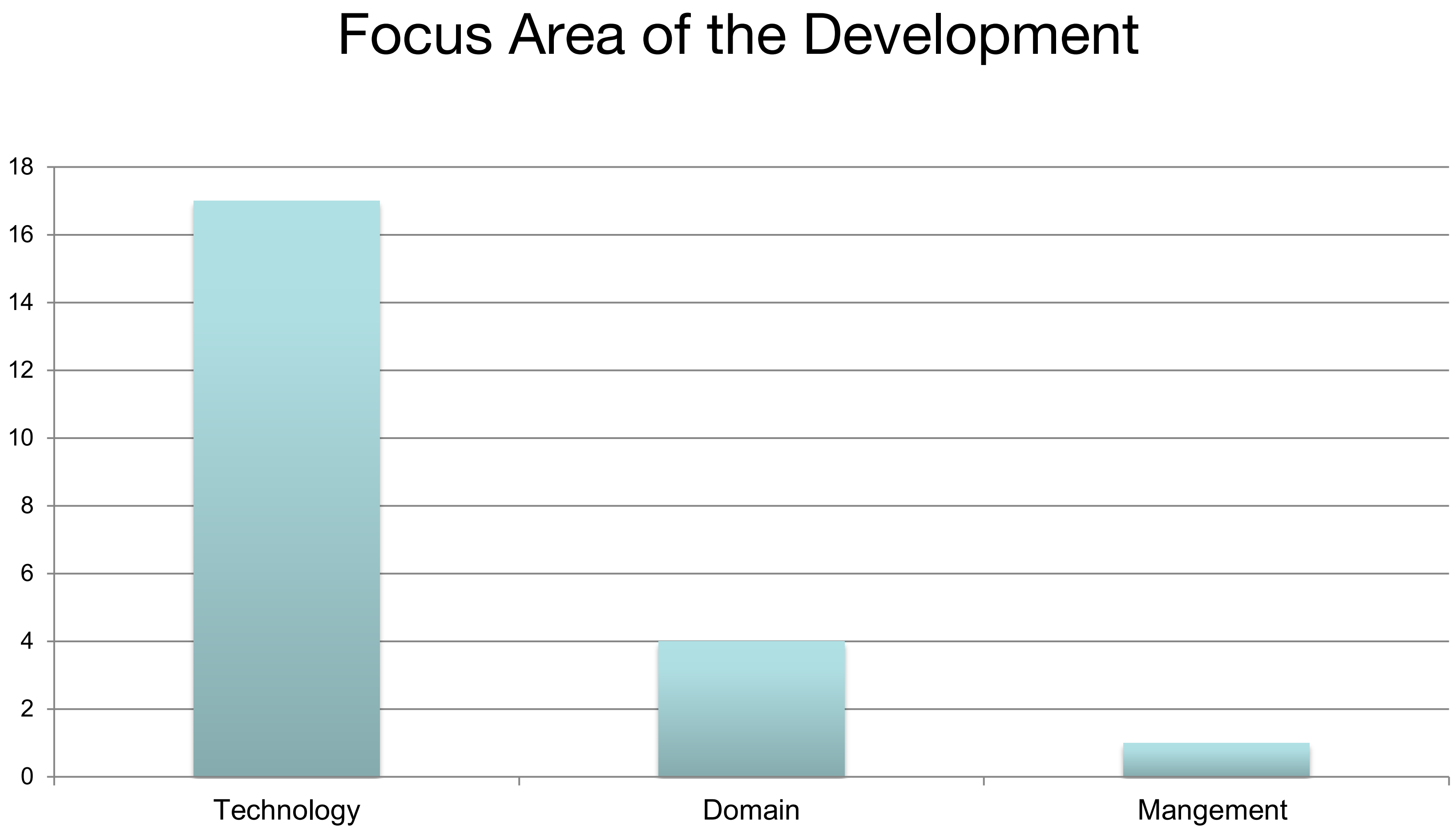
Pre-Workshop Survey - Questionnaire



Pre-Workshop Survey - Questionnaire



Pre-Workshop Survey - Questionnaire



7. Key new CI requirements and challenges in the next 5-10 years

- Data (scaling CI, performance, new technologies, e.g. ML, long-term archiving)
- Computing/networking (increasing demand, role of cloud, sensors' complexity)
- Software (long-term sustainability, reproducibility of scientific results)
- Operating and maintaining CI (configuration/management tools, SLAs, cybersecurity)
- Integration and interoperability (integration of facilities, international agreements)
- Workforce (training and retaining but also training the teachers)
- Community engagement (increasing user demand, interactions with field researchers)

Workshop Activities

Pre-

Workshop

Post-

- White papers
- Questionnaire
- Workshop Website

[Spring/Summer 2017]

- **Panels and Breakouts**
- **Key Findings**
- **Recommended Actions**

[September 6/7, 2017]

- Post-Workshop Survey
- Workshop Report
- Study of White Papers

[Fall 2017, Spring 2018]

Summary of Panels and Breakouts

- Panel 1: Integration, interoperability and reuse of CI solutions, practices
  - Facilities typically address CI challenges **independently of each other**
    - Facilities develop custom solutions in an uncoordinated manner
    - Missing opportunities to leverage existing solutions and knowledge
  - Facilities can benefit from a **trusted forum** to:
    - Facilitate discussions
    - Collect and disseminate information about addressing technical challenges, solutions, etc.
    - Provide information and potential evaluation of existing CI solutions.

    Such a forum can help:
    - Existing facilities,
    - New facilities' start up
    - Operations transferred to a new group

Summary of Panels and Breakouts

- Panel 2: Workforce development, and education and outreach
  - Significant workforce development, education, and outreach challenges while encountering **poor mission alignment to host institution** HR policies

    However, they are independently working to address these challenges
  - Fostering community learning via **increased intra-facilities communications** has the potential to form effective, network-wide workforce strategies
    - Independent programmatic successes

Summary of Panels and Breakouts

- Panel 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
  - The community suggested a topic-specific **conference focused on CI** best practices.

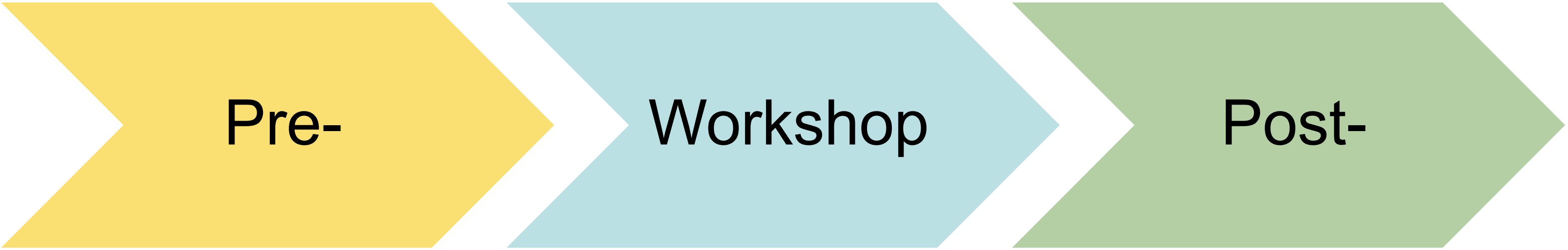
Summary of Panels and Breakouts

- Panel 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.

Recommended Actions

- Foster the creation of a **facilities' CI community** and establish mechanisms/resources to enable interaction, collaboration, and sharing
- Support the creation of a curated portal and **knowledge base** to enable the discovery and sharing of CI-related challenges, "best practices", etc.
- Establish a **center of excellence** (following a model similar to 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**
- Develop shared **metrics and methodologies** for evaluating CI
- Explore **collaborations** and synergies with facilities funded by other agencies, as well as with industry

Workshop Activities



Pre-

Workshop

Post-

- White papers
- Questionnaire
- Workshop Website

[Spring/Summer 2017]

- Panels and Breakouts
- Key Findings
- Recommended Actions

[September 6/7, 2017]

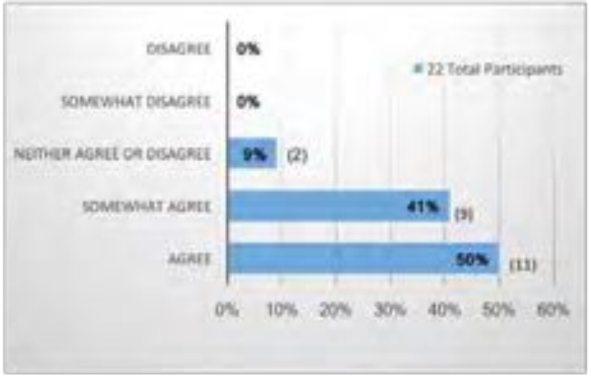
- **Post-Workshop Survey**
- **Workshop Report**
- **Study of White Papers**

[Fall 2017, Spring 2018]

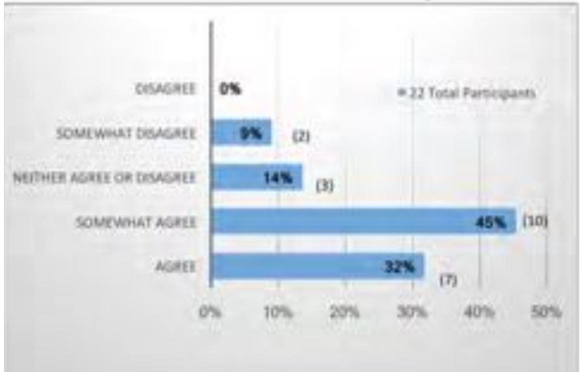
Post-Workshop Survey Questionnaire

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

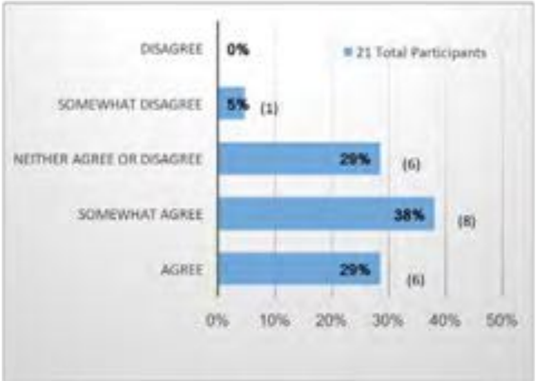
Workshop provided an understanding of current CI architecture &operations best practices at the facilities.



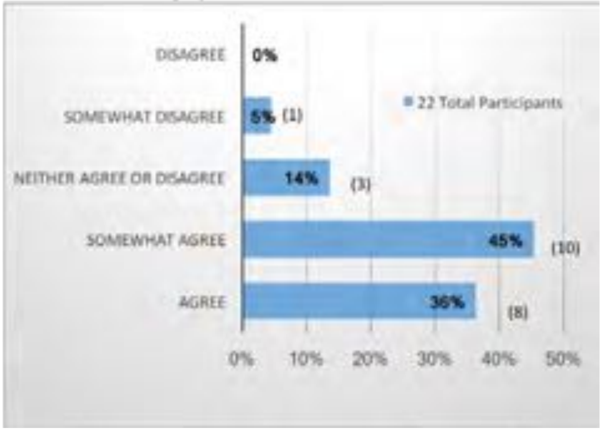
Workshop identified common requirements and solutions, as well as CI elements that can be shared across facilities.



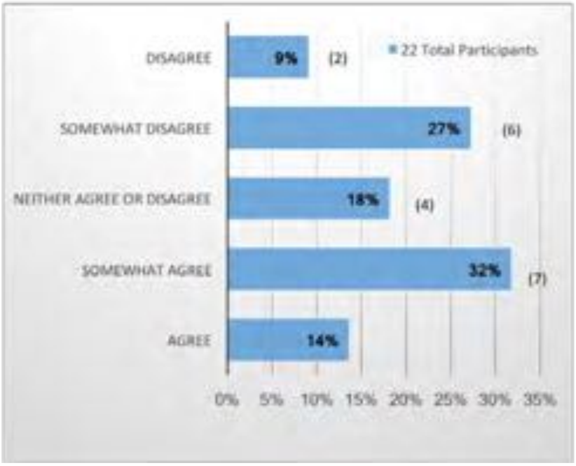
Workshop enabled CI developers to most effectively target CI needs and gaps of large facilities.



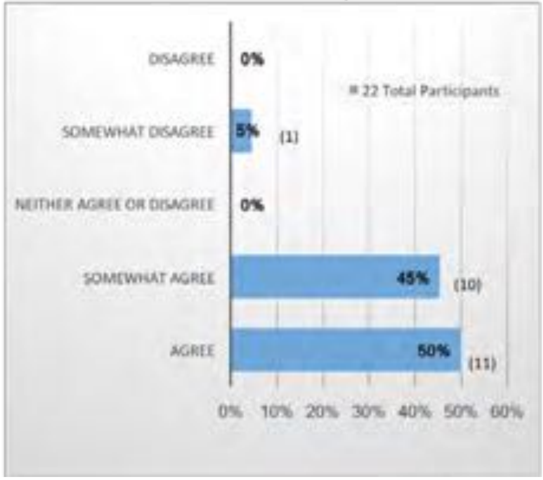
Workshop explored opportunities for interoperability between the large facilities.



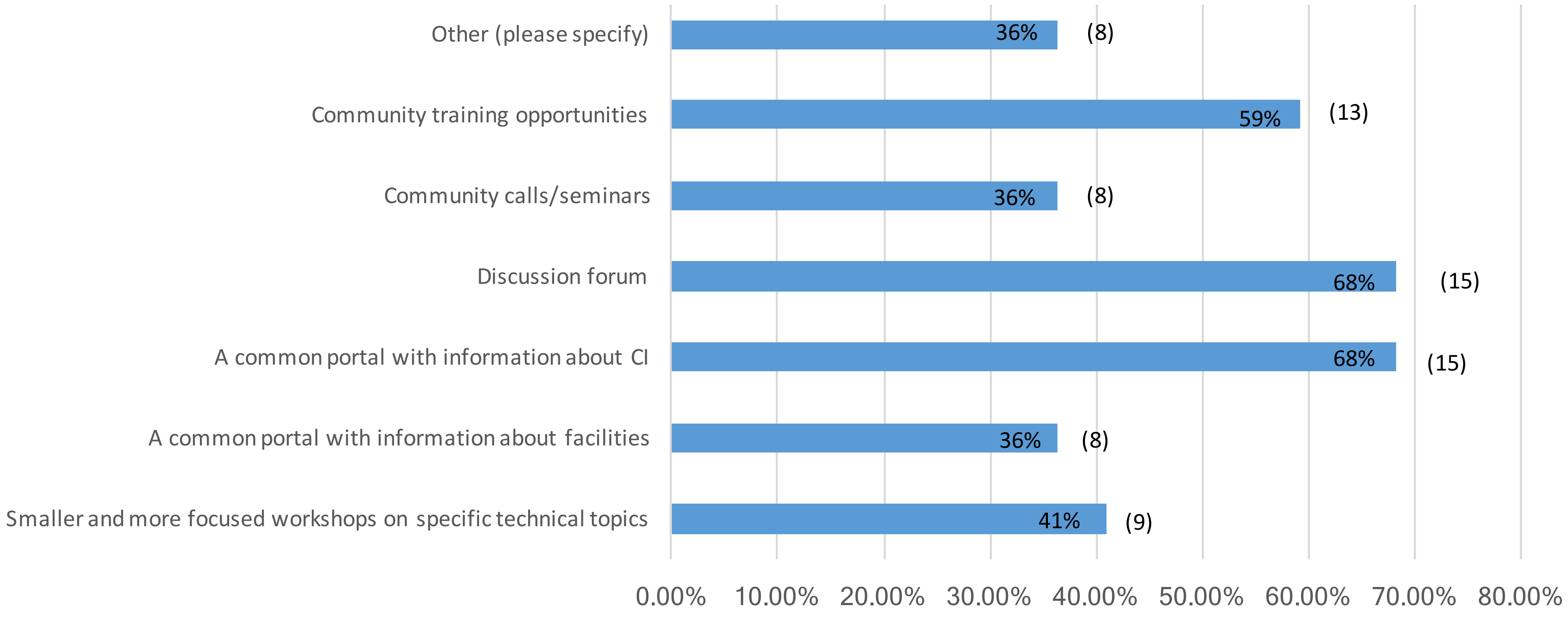
The workshop developed guidelines, mechanisms, and processes that can assist future large facilities in constructing and sustaining their CI.



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



Should this workshop be held again? If so, what should be the focus?

- Focus on **collaboration opportunities**, interoperability across facilities
- Finding specific areas of CI overlap among facilities and **forming partnership**
- **Polling** the community and sharing best practices
- NSF to develop **communication channels**, or form an organization to support intra-LSF collaboration
- Working to provide a **sustainable home** for cyberinfrastructure resources
- Add some centered discussion from the **point of view of CI users**
- Etc.

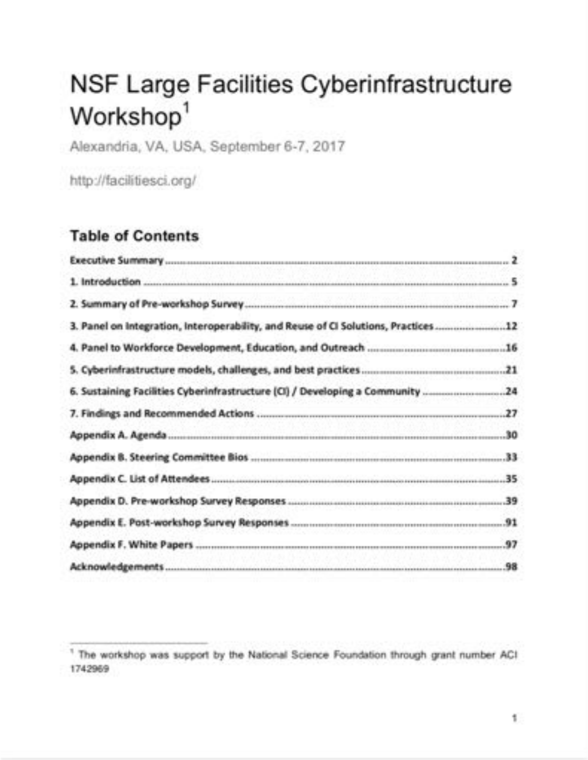
Workshop Report

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

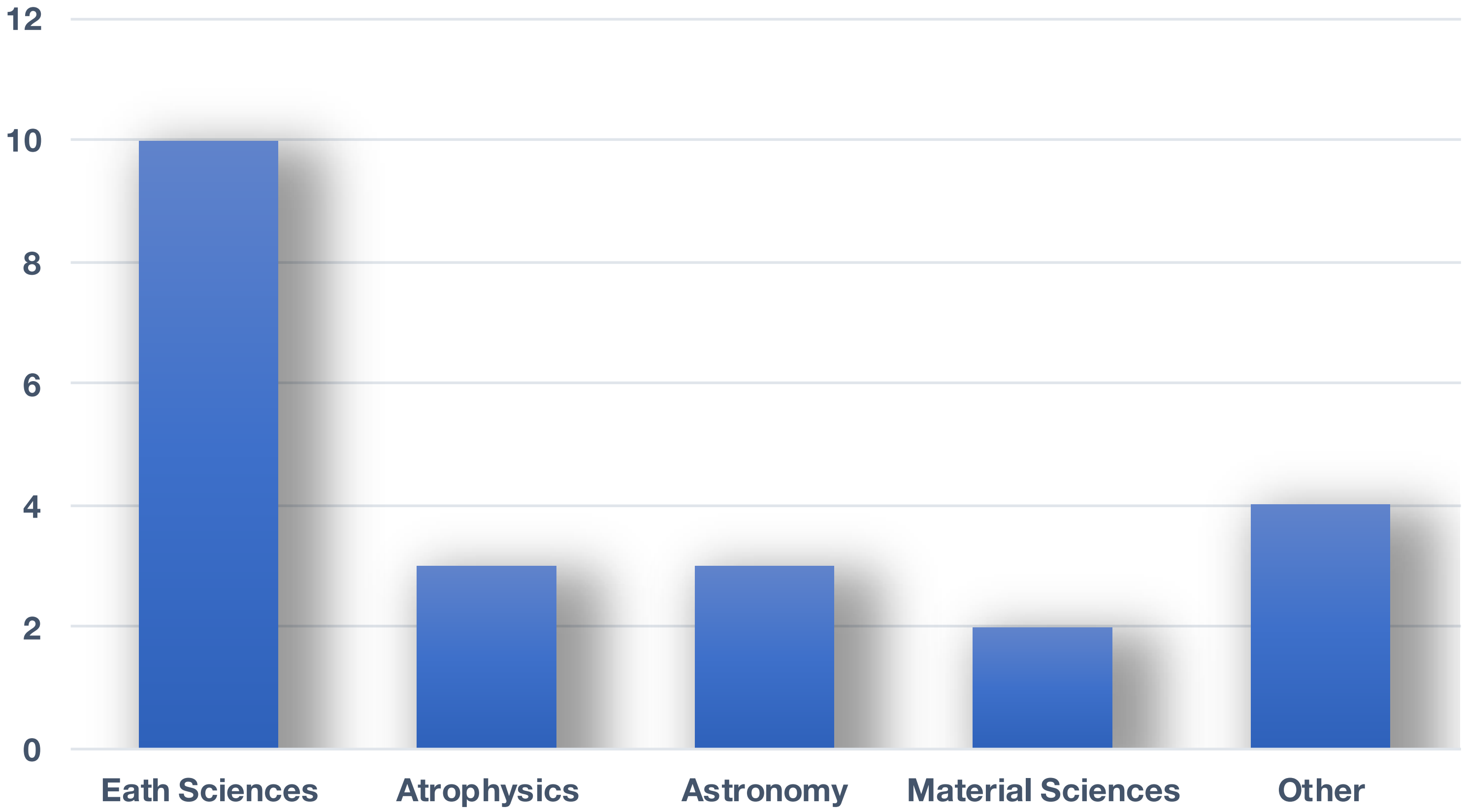
1

Report available at ***facilitiesci.org***

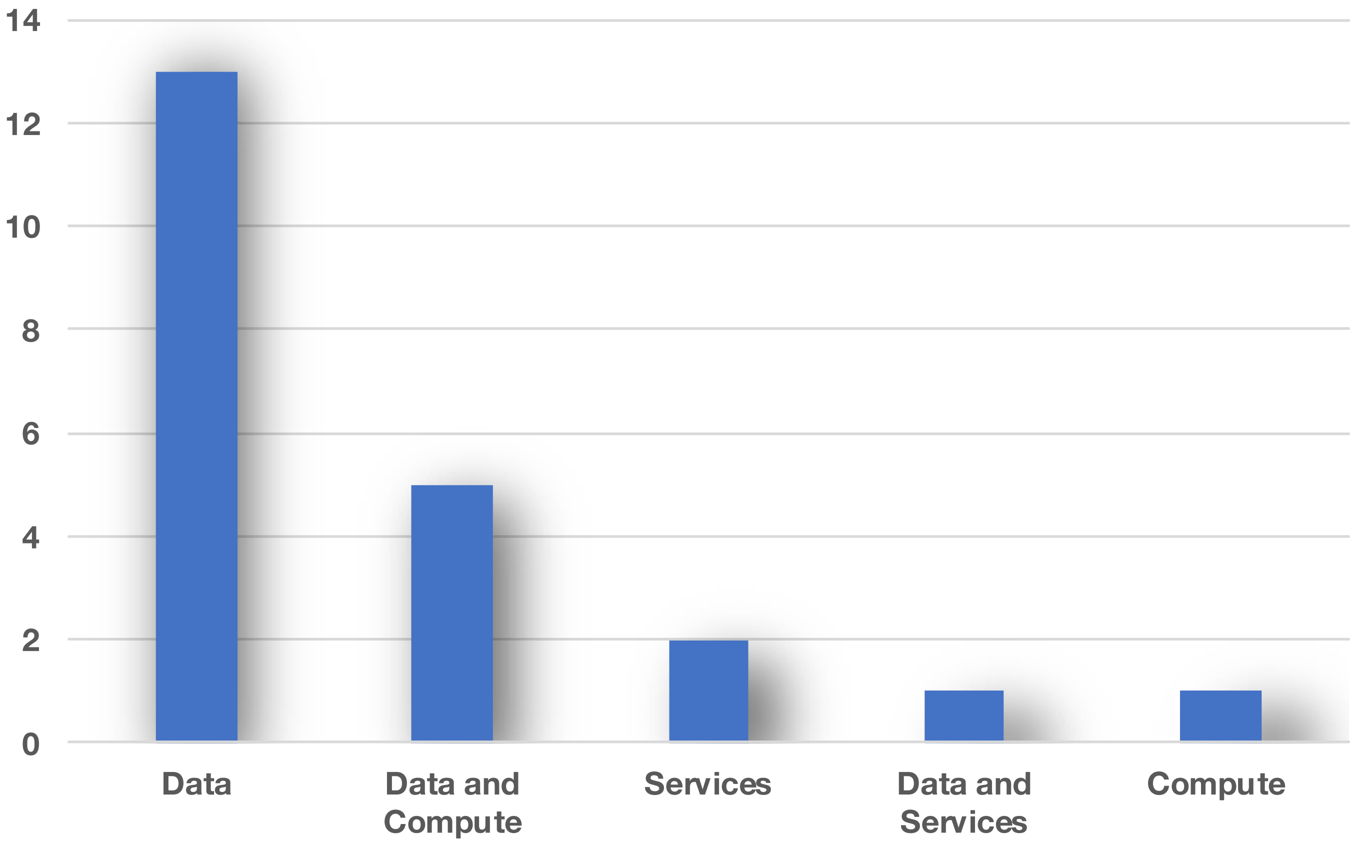
Study of White Papers

- Key Attributes (categorized)
  - **Area/Domain**
  - **Main types of offerings**
  - **Data mechanisms and types**
  - **Data storage mechanisms**
  - **Resiliency mechanisms**
  - **Data delivery mechanisms**
  - **Access model**
- Other Attributes/Properties
  - System architecture (e.g., bare metal cluster vs. virtualized, information vs. information lifecycle management system)
  - Data processing capabilities
  - Deployment model (e.g., on-premises vs. offsite [e.g., Cloud])
  - Operation model
  - Cyber-security mechanisms
  - Size
  - Age

Breakout by Area/Domain

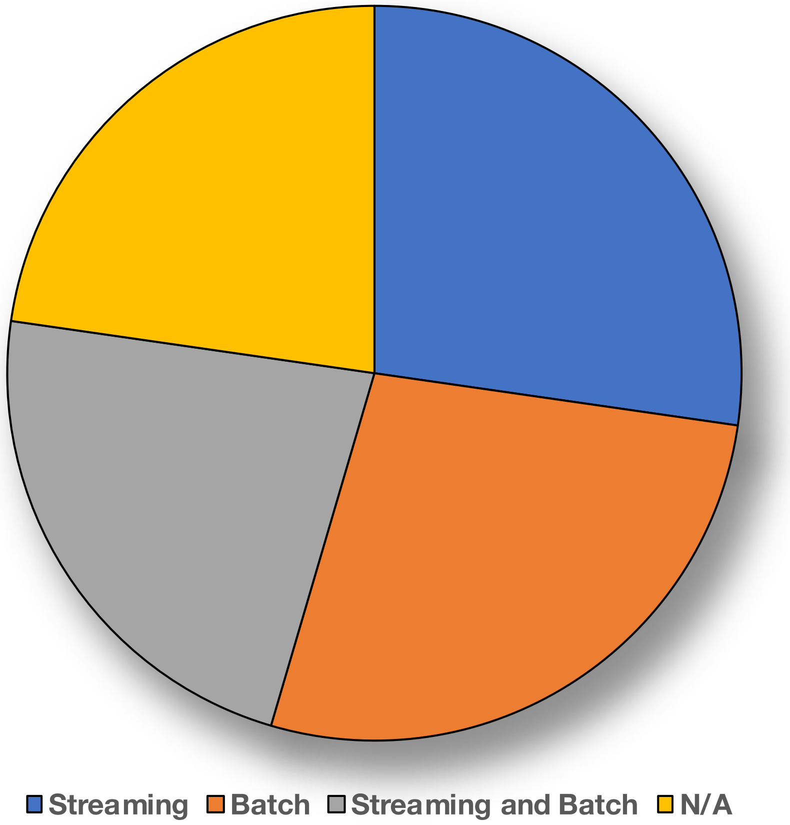


Main Type of Offerings

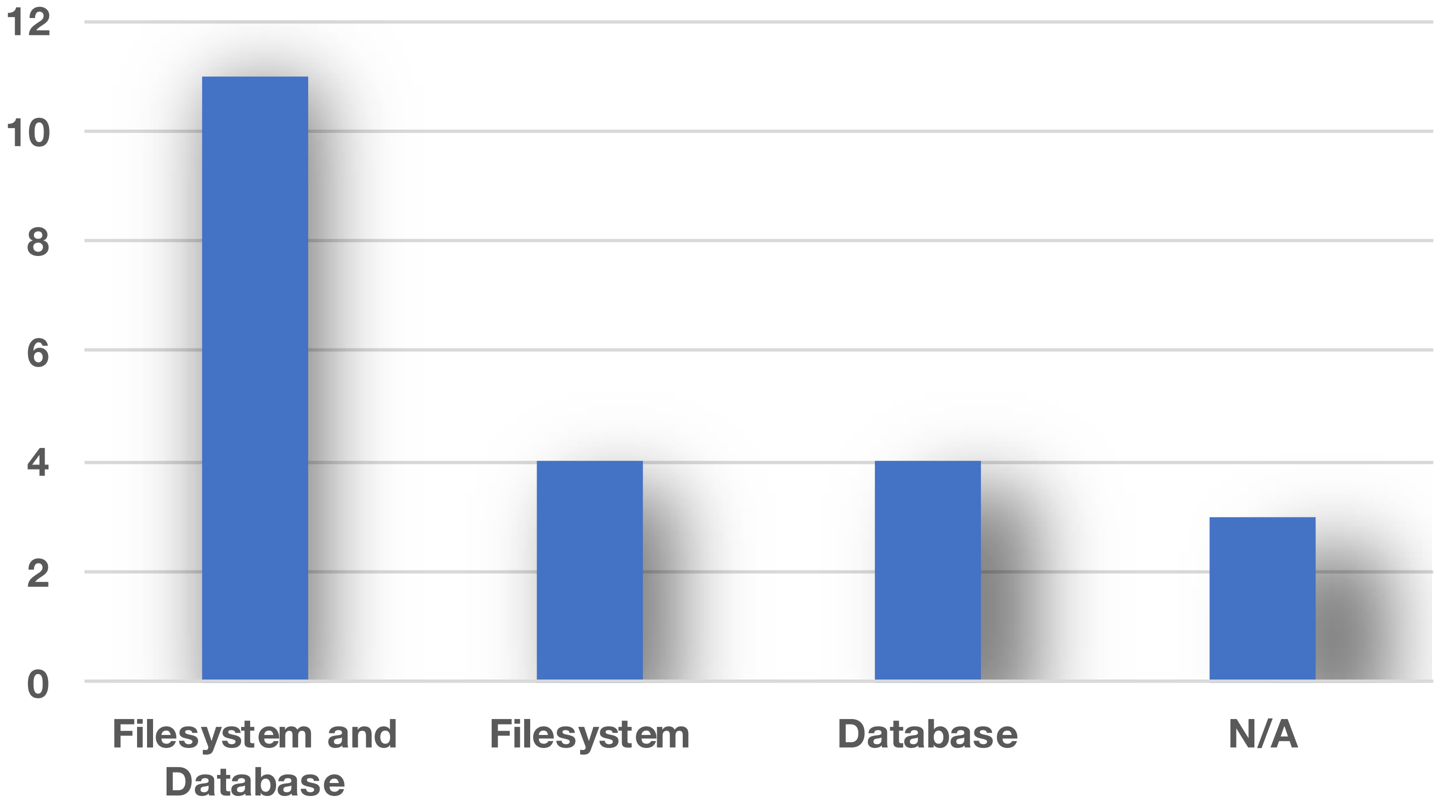


Another perspective: Instrument focused vs. compute/service-centric

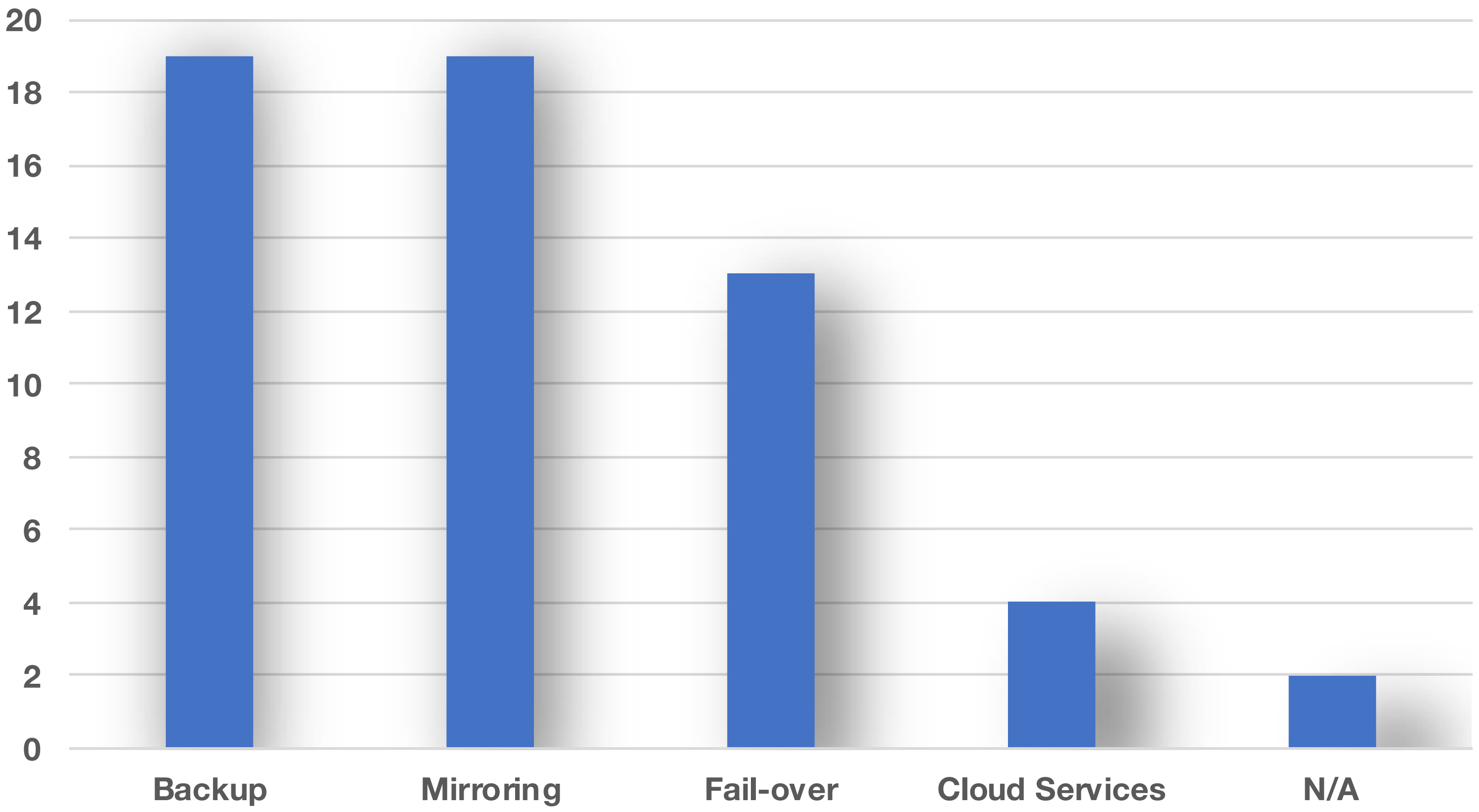
Data Collection Mechanisms and Types



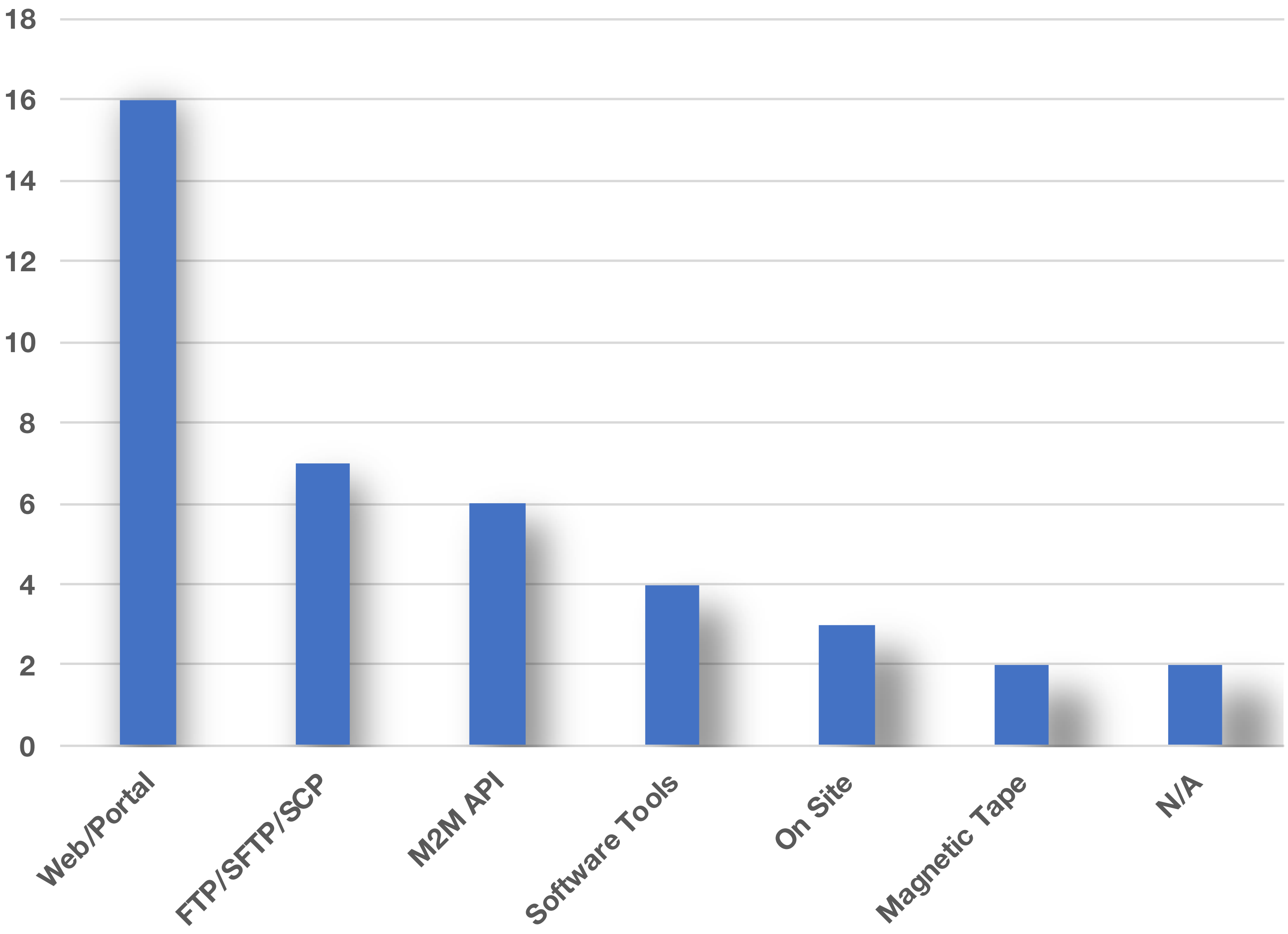
Data Storage Mechanisms



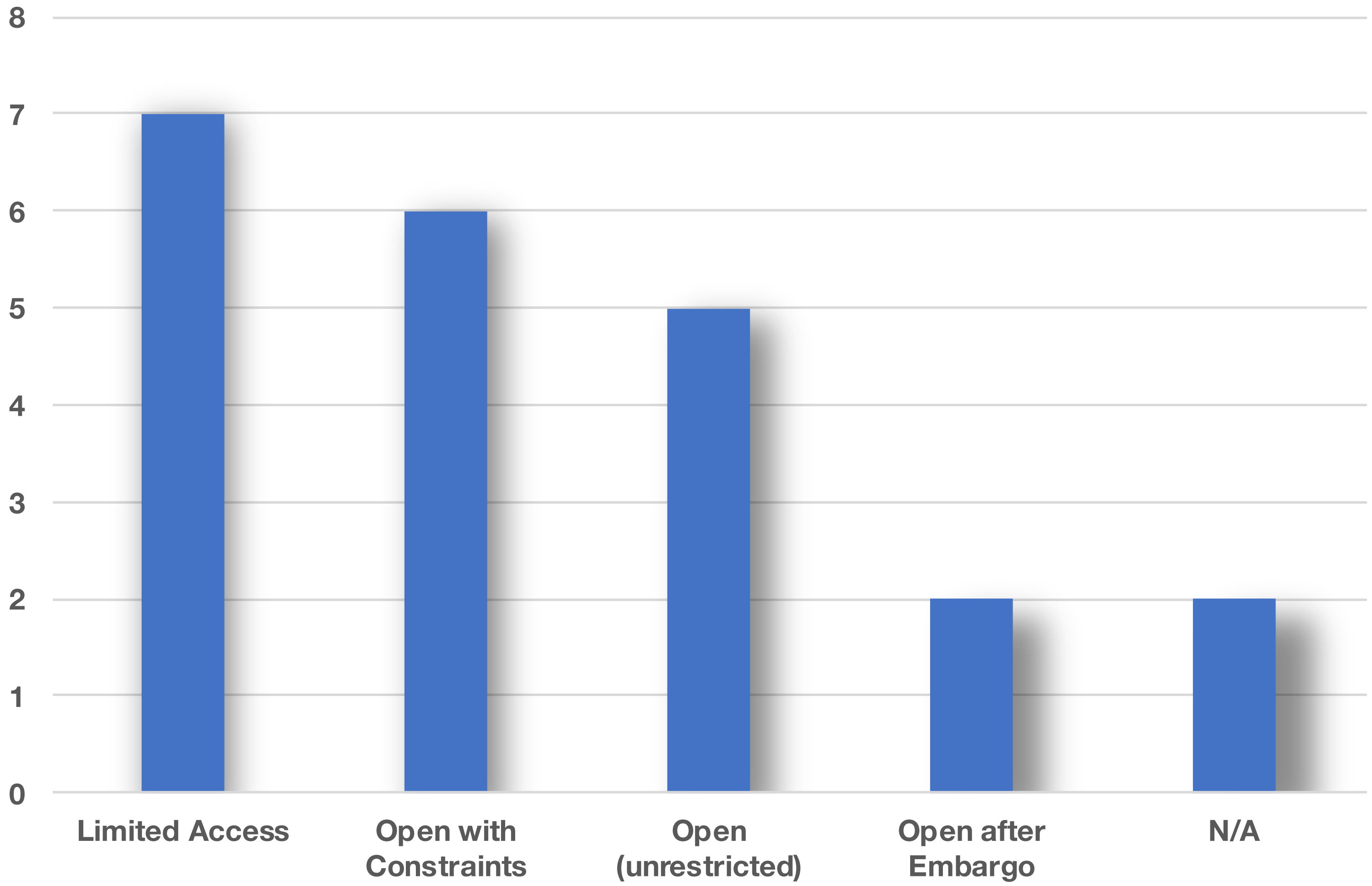
Resiliency Mechanisms



Data Delivery Mechanisms



Access Model



Challenges – Looking Forward

- Cyber-security
- Need for new data delivery models (e.g., real-time data)
- Disconnect between Large Facilities and existing Cyberinfrastructure for supporting data-driven workflows
- Cyberinfrastructure interoperability and interoperation
- Ongoing efforts (use case)

Cyber-security

- “Traditional” mechanisms
  - Encryption
  - Virtual Private Networks
  - Two-factor authentication
  - Federated Identify Management, etc.
- Cyber-security programs/frameworks
  - NSF CTSC
  - NIST 800-53, FISMA, HIPAA, ISO 27000, etc.
- Specific risks for facilities
  - National security issues
  - Issues related to physical access to instrumentation, etc.
- New paradigms
  - Ransomware
  - Gaining access to compute resources for virtual currency mining, etc.

Need for Real-time Data Delivery

- Data is increasing in scale, heterogeneity, and richness
  - Data downloads and local processing are no longer feasible
- Integrating observatory data into scientific workflows is a growing challenge
  - **New delivery modes** for data and data-products are **essential!**
- The CI must explore richer and **more intelligent** data delivery mechanisms
  - Leverage **machine learning** techniques to stream the right data to the users at the right time (early experiences next)

Leverage ACI for Supporting Data-driven Workflows

- Enable workflows that when triggered can seamlessly orchestrate the entire data-to-discovery pipeline
  - Provide distributed content-delivery networks (CDNs)
  - Manual query and processing of data (at NSF-ACI)
  - Data-driven query and processing of data
  - Data-driven query, aggregation, and processing across multiple data-stores
- Leverage existing high-speed interconnects (internet2)

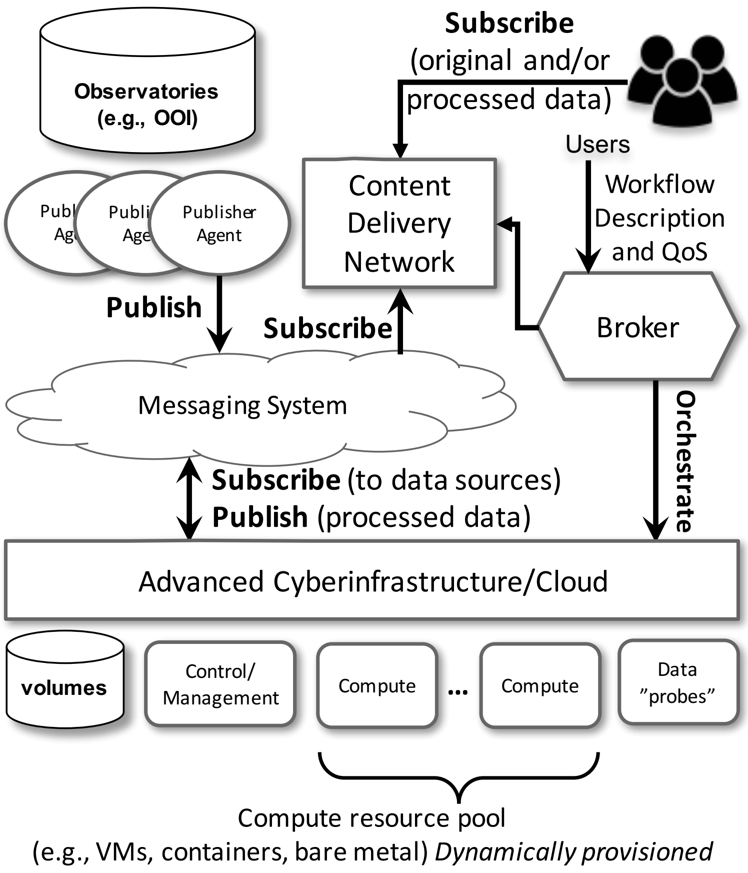
Submarine: Re-thinking Data Delivery

- **Ongoing efforts** include prototyping scenarios based on publish/subscribe-based streaming and considering approximation techniques
- Distributed content-delivery networks (CDNs)

  Data-push, publish/subscribe/notify semantics for data and data products
  - Leverage "content-push" revolution in social media
- Data-driven (content, location, quality) workflows that seamlessly orchestrate the entire data-to-discovery pipeline

Digital Still Camera – Image Analysis on CI/Cloud

Architecture



Observatories (e.g., OOI)

Publisher Agent

Publisher Agent

Publisher Agent

Publish

Subscribe (original and/or processed data)

Users

Content Delivery Network

Workflow Description and QoS

Subscribe

Broker

Messaging System

Orchestrate

Subscribe (to data sources)
Publish (processed data)

Advanced Cyberinfrastructure/Cloud

volumes

Control/ Management

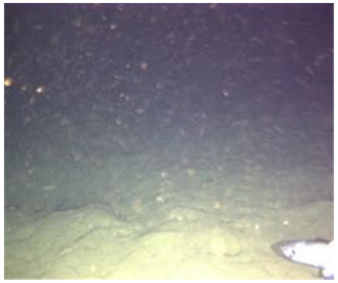
Compute

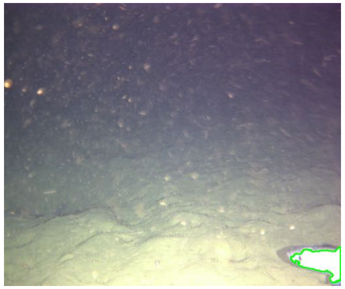
...

Compute

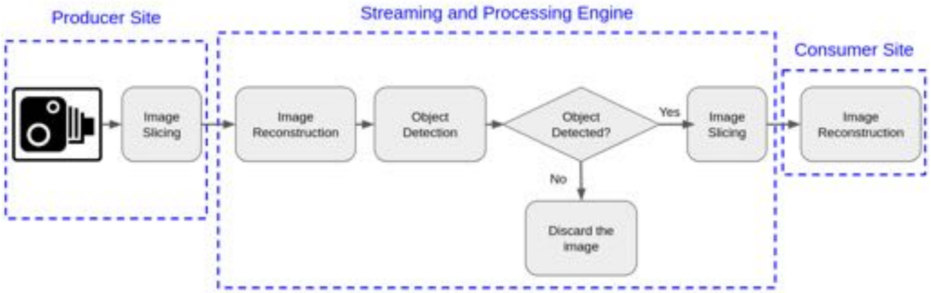
Data "probes"

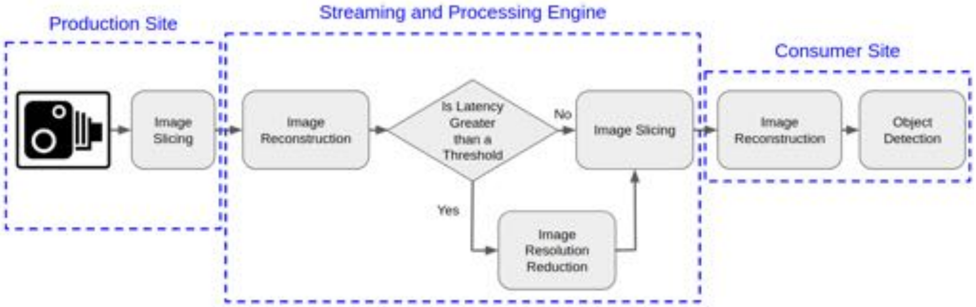
Compute resource pool
(e.g., VMs, containers, bare metal) *Dynamically provisioned*





Implementations





“Take Home” Messages

- **Need for establishing facilities’ CI community** and mechanisms/resources to enable the community to interact
  - Not only better efficiencies by coordinating large facilities’ CI efforts, but also potential for more advanced insights (e.g., data-driven workflows combining different sources)
- Need for **re-using existing ACI investments** and experiences
  - Not only compute/data platforms but also software distribution, software management and sustainability, etc.
  - Interoperation and interoperability
- Need for “CI best practices” and **trusted entity** (e.g., center for excellence), accepted by the academic community
  - Evolving technologies and facilities’ requirements
  - Workforce: training on these practices
  - Science community perception matters (e.g., engagements: what vs. how)

Thank You!

Ivan Rodero
Rutgers Discovery Informatics Institute (RDI^2)

irodero@rutgers.edu

facebook.com/rutgersRDI2

@RDI2_rutgers

rdi2.rutgers.edu