2019 NSF Large Facilities Cyberinfrastructure Workshop

Connecting Large Facilities and Cyberinfrastructure

September 16-17, 2019

http://facilitiesci.org

WIFI:

Login: **Facilities CI**
Password: **meeting123**

Funded by the National Science Foundation under award #1933353, William Miller, OAC

2019 NSF Large Facilities Cyberinfrastructure Workshop

Connecting Large Facilities and Cyberinfrastructure

September 16-17, 2019

http://facilitiesci.org

Ewa Deelman, University of Southern California

Funded by the National Science Foundation under award #1933353, William Miller, OAC

CI CoE PILOT

Steering Committee

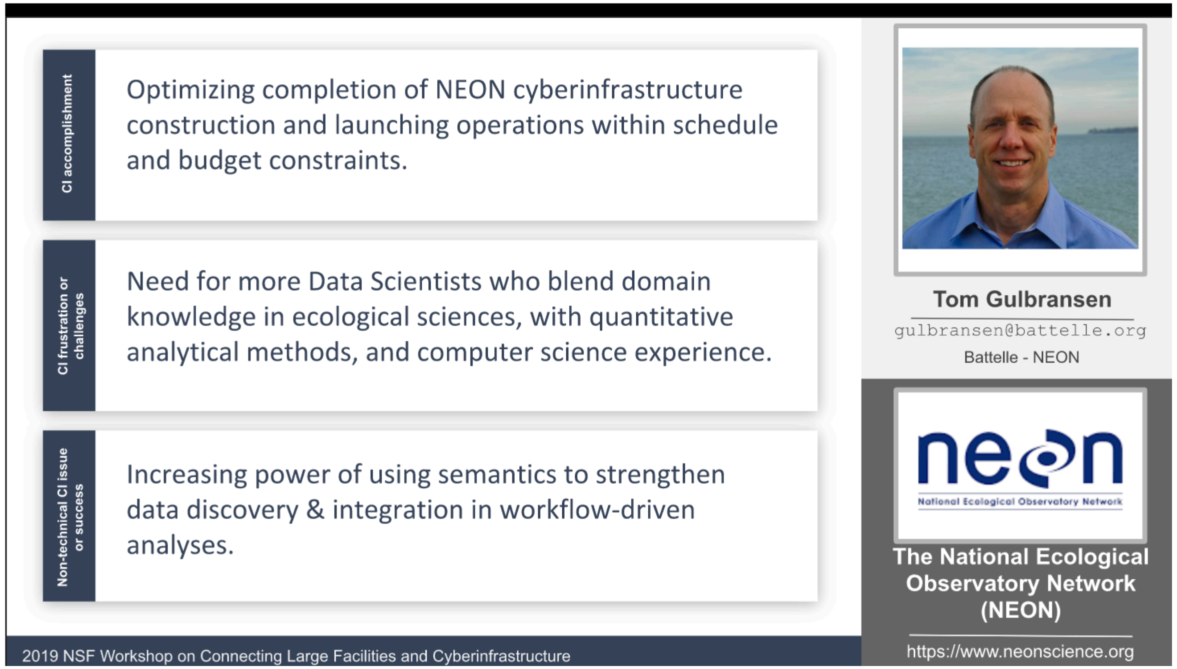
	Adam Bolton	National Optical Astronomy Observatory		**Kate Keahey**	Chameleon, Argonne National Laboratory
	Brian Bockelman	OSG, Morgridge Institute		**Marina Kogan**	University of Utah
	Tom Cheatham	CHPC, University of Utah		**Dan Stanzione**	Texas Advanced Computing Center
	Tom Gulbransen	NEON, Battelle		**Daryl Swensen**	Regional Class Research Vessel, U. of Oregon

School of Engineering
Information Sciences Institute

Organizers: Rafael Ferreira da Silva, Jasmine Mann, Mats Rynge, USC

- Pre-workshop survey (April/May) (**43**)
- Workshop participant Survey (**27**)
- CI Practitioner Survey (**49**)

- CI Calling Cards (**51**!):
  - Biggest CI accomplishment,
  - Biggest CI frustration or challenge
  - Non-technical frustration or accomplishment when building CI
  - **You can still add your own**
  - **We will make them searchable and expand**



CI accomplishment

Optimizing completion of NEON cyberinfrastructure construction and launching operations within schedule and budget constraints.

CI frustration or challenges

Need for more Data Scientists who blend domain knowledge in ecological sciences, with quantitative analytical methods, and computer science experience.

Non-technical CI issue or success

Increasing power of using semantics to strengthen data discovery & integration in workflow-driven analyses.

Tom Gulbransen
gulbransen@battelle.org
Battelle - NEON

neon
National Ecological Observatory Network

The National Ecological Observatory Network (NEON)

https://www.neonscience.org

2019 NSF Workshop on Connecting Large Facilities and Cyberinfrastructure

CI CoE PILOT

Workshop Goals

Theme:
Connecting Large Facilities, Connecting CI, Connecting People

Cyberinfrastructure "consists of computing systems, data storage systems, advanced instruments and data repositories, visualization environments, and people, all linked together by software and high performance networks to improve research productivity and enable breakthroughs not otherwise possible." [1]

Workshop Goal: Foster discussions and collaborations amongst NSF-funded Large Facilities and CI projects

USC Viterbi School of Engineering *Information Sciences Inst*

[1] M. Parashar, S. Anderson, E. Deelman, V. Pascucci, D. Petravick, and E. M. Rathje, "2017 NSF Large Facilities Cyberinfrastructure Workshop," 2017. [Online]. Available: http://facilitiesci.org/assets/reports/facilitiesci-workshop-report-11-17.pdf

Workshop Topics

Technical:

- What are the CI challenges that need to be addressed to support LF science?
- Where does LF CI end and the user CI begin (issues of data sharing, reproducibility)?
- Can we better utilize current CI investments?
- What are the opportunities to share CI services?

Socio-technical:

- What are the opportunities for collaboration amongst LFs and other Large CI projects?
- What are the non-technical issues that influence CI development and how they can be collaboratively addressed?
- Enhancing the CI workforce: what are the challenges and solutions?
- How can we build a CI community: what are the impediments and opportunities?

Previous Workshops

Manish Parashar (PI and Chair), Rutgers University and OOI
Stuart Anderson, LIGO
Ewa Deelman, USC
Valerio Pascucci, University of Utah
Donald Petravick, LSST
Ellen M. Rathje, NHERI

NSF Large Facilities Cyberinfrastructure Workshop

IceCube

September 2017 Workshop report at http://facilitiesci.org/

- 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 Key Findings

- The need for, and benefits of, **close interactions, collaborations, and sharing** among the facilities and with the CI communities: sharing of CI related **expertise, technical solutions, best practices, and innovations** across NSF large facilities as well as DOE, NIH, NASA,
- 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
- **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.
- **Scientific Cyberinfrastructure: Cybersecurity (Center for Trustworthy Scientific Cyberinfrastructure– now Trusted CI)** was explicitly and repeatedly noted as an effective model that should be explored to address this gap.

Workshop Key Recommendations

- **Establish a center of excellence** (following a model similar to the NSF-funded Trusted CI) as a resource providing expertise in CI technologies and effective practices related to large-scale facilities as they conceptualize, start up, and operate.
- 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 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.

Questions We Posed

- Are we ready to build a CI community?
- How do we build a CI community?
- How do we enhance collaborations across large facilities and CI projects?
- How do we capture knowledge, effective practices in a way that is relevant, evolving, and impactful?
- How do we maintain and enhance/increase the CI talent pool?

CI CoE PILOT

Pilot Study for a Cyberinfrastructure Center of Excellence

Funded by the National Science Foundation
Grant #1842042

Ewa Deelman, USC (PI)

Co-PIs:
Anirban Mandal, RENCI

Jarek Nabrzyski, Notre Dame University

Valerio Pascucci and **Rob Ricci**, University of Utah

Funded by NSF

Co-funded by BIO (Roland Roberts) and OAC (William Miller)
10/2018- 9/2020

USC Viterbi
School of Engineering
Information Sciences Institute

UNIVERSITY OF NOTRE DAME

THE UNIVERSITY OF UTAH

CENTER FOR APPLIED CYBERSECURITY RESEARCH

renci

Project Goals

Award #1842042

Develop a model and a plan for a Cyberinfrastructure Center of Excellence

- Platform for knowledge sharing and community building
- Key partner for the establishment and improvement of Large Facilities with advanced CI architecture designs
- Grounded in re-use of dependable CI tools and solutions
- Forum for discussions about CI sustainability and workforce development and training
- Pilot a study for a CI CoE through close engagement with NEON and further engagement with other LFs and large CI projects.

10/2018– 9/2020

Overall Strategy

1. Recognize the expertise, experience, and mission-focus of Large Facilities
2. Engage with and learn from current LFs CI
3. Build on existing knowledge, tools, community efforts
   -Avoid duplication, seek providing added value,
4. Prototype solutions that can enhance particular LF's CI
   -Keep a separation between our efforts and the LF's CI developments
5. Build expertise, not software
6. Work with the LFs and the CI community on a blueprint for the CI CoE

Build partnerships:

- Trusted CI (identity management): share personnel
- Open Science Grid (data and workload management): share expertise
- Campus Research Computing Consortium (CaRCC): workforce development

Project Team

USC

Ewa Deelman
Mats Rynge
Karan Vahi Loïc Pottier
Rafael Ferreira da Silva
Ryan Mitchell

Automation, Resource Management, Workflows

RENCI

Anirban Mandal
Ilya Baldin
Laura Christopherson
Paul Ruth
Erik Scott

Resource Management, Networking, Clouds

Project Collaborators

University of Notre Dame

Jarek Nabrzyski
Jane Wyngaard
Charles Vardeman

Workforce development, Sensors, Semantic technologies

University of Utah

Valerio Pascucci, Rob Ricci,
Timo Bremer, Attila Gyulassy,
Steve Petruzza

Data management, visualization, clouds, large-scale CI deployment

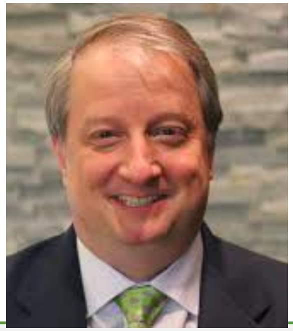
Indiana University

Susan Sons (co-funded by Trusted CI)
Von Welch (unfunded collaborator)

Cybersecurity

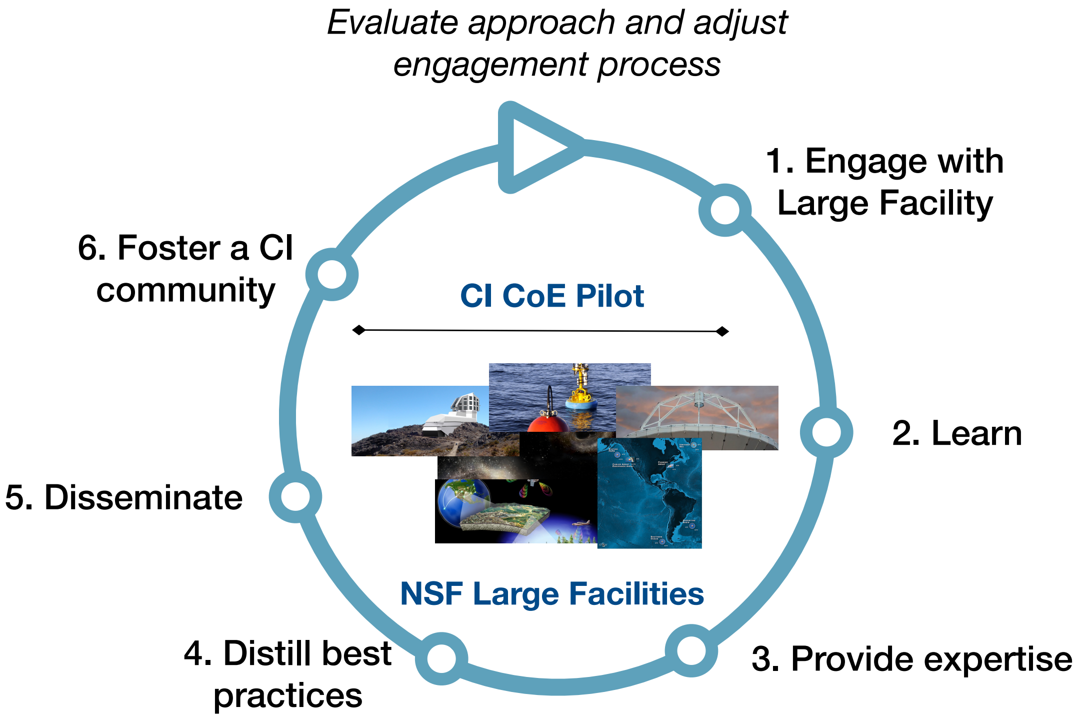
Advisory Board

- **Stuart Anderson**, Caltech
- **Pete Beckman**, ANL, Northwestern University
- **Tom Gulbransen**, Battelle
- **Bonnie Hurwitz**, University of Arizona
- **Miron Livny**, University of Wisconsin, Madison
- **Ellen Rathje**, University of Texas at Austin
- **Von Welch**, Indiana University
- **Michael Zentner**, Purdue University



Engagement Methodology

Developing and improving Engagement Model


Evaluate approach and adjust engagement process
1. Engage with Large Facility
2. Learn
3. Provide expertise
4. Distill best practices
5. Disseminate
6. Foster a CI community
CI CoE Pilot
NSF Large Facilities

Process for Engagement with a Facility

- Engage at the management level, potentially seek introductions from NSF PO, participate in meeting (LF Workshop, LF CI Workshop)
- Initial virtual technical group discussions to define possible avenues of engagement
- In person meeting with a number of technical personnel
- Identity topics for engagement
- Set up working groups
- Follow up email and conference call discussions focused on particular topics/working groups
- Bigger group discussions/checkpointing
- Reports of engagement, gather feedback from the project engaged

NEON Engagement, October 2018-

- Engagement facilitated by NSF
- Engagement Goals:
  - Increase Pilot’s understanding of NEON’s cyberinfrastructure architecture and operations
  - Increase NEON’s understanding of the Pilot’s goals and expertise
  - Select & scope mutually beneficial opportunities to prototype or learn from CI methods
- Engagement Process
  - In-person management meeting
  - NEON shared a number of design documents
  - Team conference calls
  - Meeting with NEON
    - November 2018: Identified topics and formed working groups
    - August 2019: took stock, summarized

CI CoE PILOT

Working Groups and Products co-lead by the Pilot and NEON

NSF

Working group	Goals	Products
Data Capture	Develop demonstrators and comparisons of the multiple architectures for data capture at the sensor to data deposition in a repository	• **Prototype**: architecture demo on github: https://github.com/cicoe/SensorThingsGost-Balena
Data Life Cycle & Disaster Recovery	Develop a general set of DR requirements and policies that can inform the LFs about best practices for DR and how those can be adapted for specific facilities.	• **Document:** Disaster recovery template • **Document:** Filled out template example (IceCube) • **Webinar**: Best Practices for NSF Large Facilities: Data Life Cycle and Disaster Recovery Planning
Data Processing	Provide support and distill best practices for workflows and services related to the processing of data.	• **Paper:** "Exploration of Workflow Management Systems Emerging Features from Users Perspectives" (Submitted to a SC'19 workshop)
Data Storage, Curation, & Preservation	Compare and be able to consult on different data storage, curation and preservation technologies.	• **Document**: Competency questions based on scenarios that domain experts may use Google dataset search for NEON dataset discovery • **Presentation**: at ESIP on schema.org • Small containerized **prototype** of publishing neon vocabularies as linked data and linked data connection

Working Groups and Products

Working group	Goals	Products
Data Visualization & Dissemination	Understand the access, visualization and user interaction workflows in large facilities. Distill best practices and provide solutions to improve the access and usability of the available data.	• **Document** describing AOP data visualization cyberinfrastructure • **Online demo and video**: Visualizing AOP Data-- https://cert-data.neonscience.org/data-products/DP3.30010.001
Identity Management	Understand current practice in authentication and authorization and help mature practice across the NSF Large Facilities.	• **Production deployment**: Connection to CI Logon NEON data download (using existing university / organization credentials) https://cert-data.neonscience.org/home • **Paper:** NEON IdM Experiences (in submission to NSF Cybersecurity Summit)
Engagement with Large Facilities	Engage with Large Facilities and other large cyberinfrastructure projects to foster knowledge and effective practice sharing; 2) define avenues of engagement, modes of engagement, and plan community activities.	• **Document**: LF engagement template • **Presentations:** SCIMMA project meeting, 2019 LF meeting, PEARC'19 • **Paper:** Invited e-Science 2019 paper

Contact: Ewa Deelman, deelman@isi.edu

Engagement with Large Facilities Lessons Learned

1. Importance of f2f discussions, building relationships and trust
2. Benefits of formalizing the engagement: expectation, timelines, resources to use
3. Importance of LF priorities and challenges, importance of good timing
4. Organizing work around working groups and work products
5. Be open to learn about what works, don't fix it (workflow management)
6. Co-existence of old and new systems, making for a heterogeneous CI landscape

Tom Gulbransen

NEON's perspective

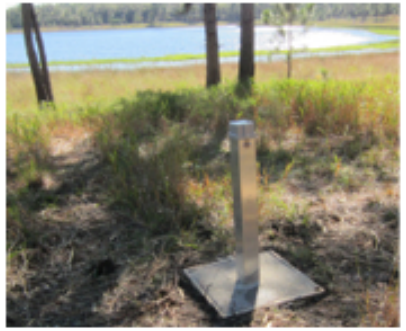
Project Manager for Cyber Infrastructure and Data Products Development

National Ecological Observatory Network Mission

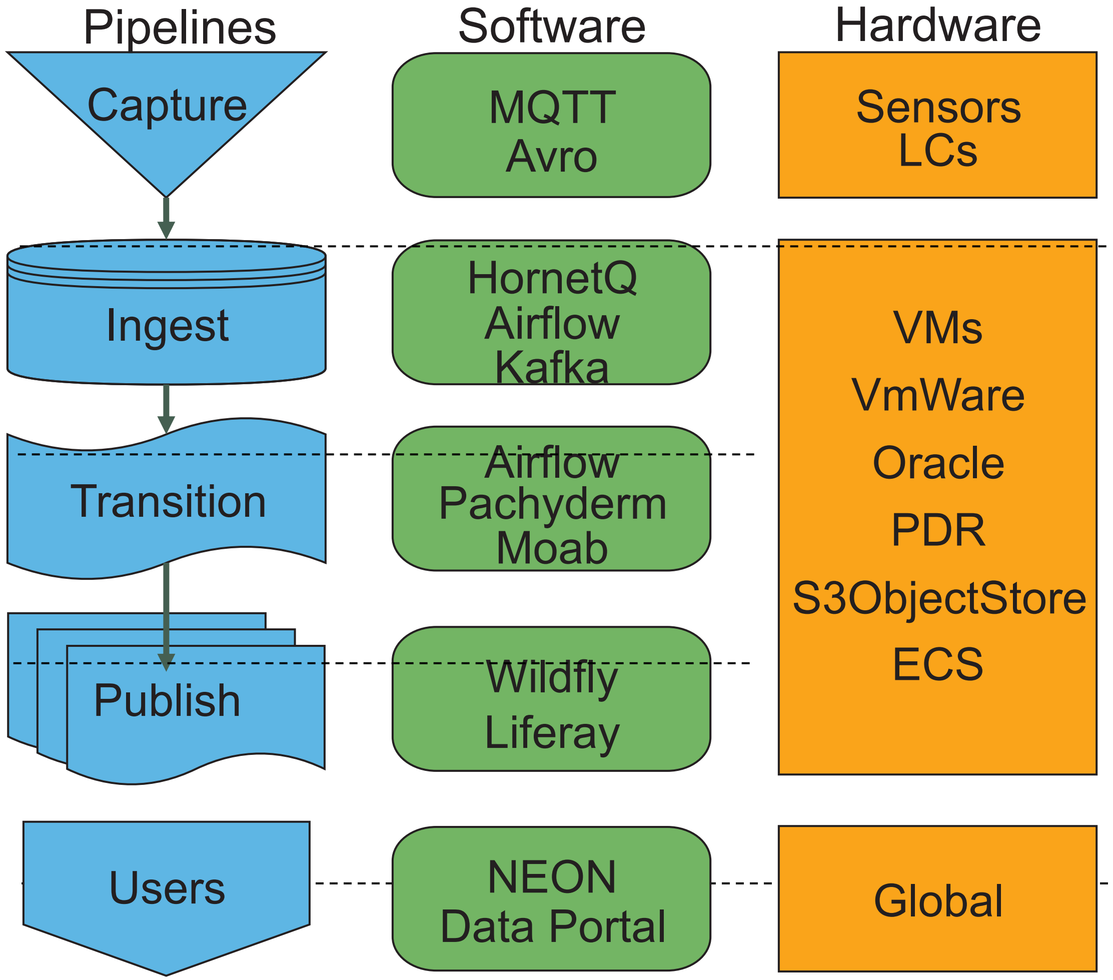
NEON provides a coordinated national system for monitoring critical ecological and environmental properties at multiple spatial and temporal scales.

…transformative science

…workforce development



NEON Cyberinfrastructure Overview

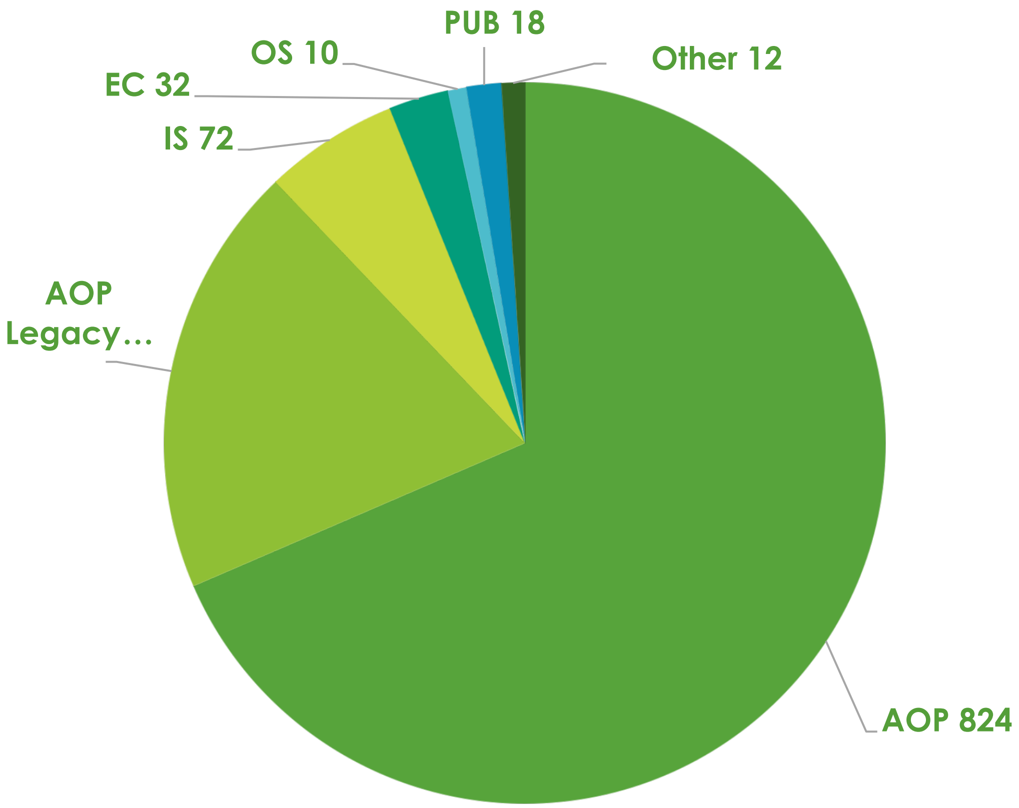


Pressure to be more effective, efficient, sustainable, & scalable

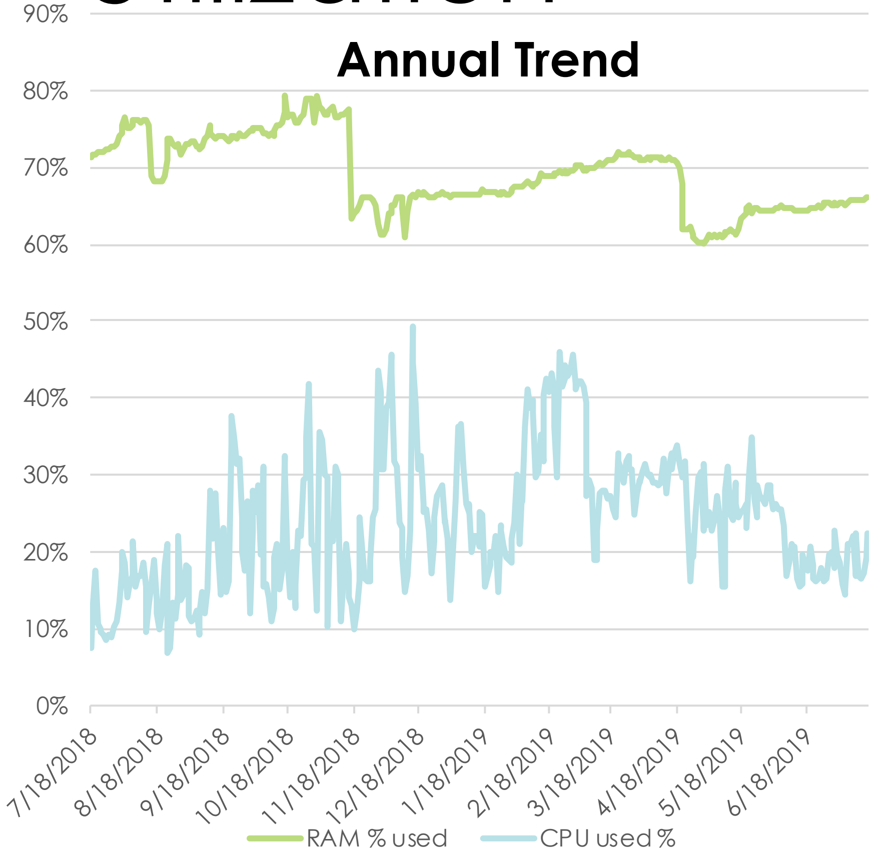
NEON CI Storage Utilization

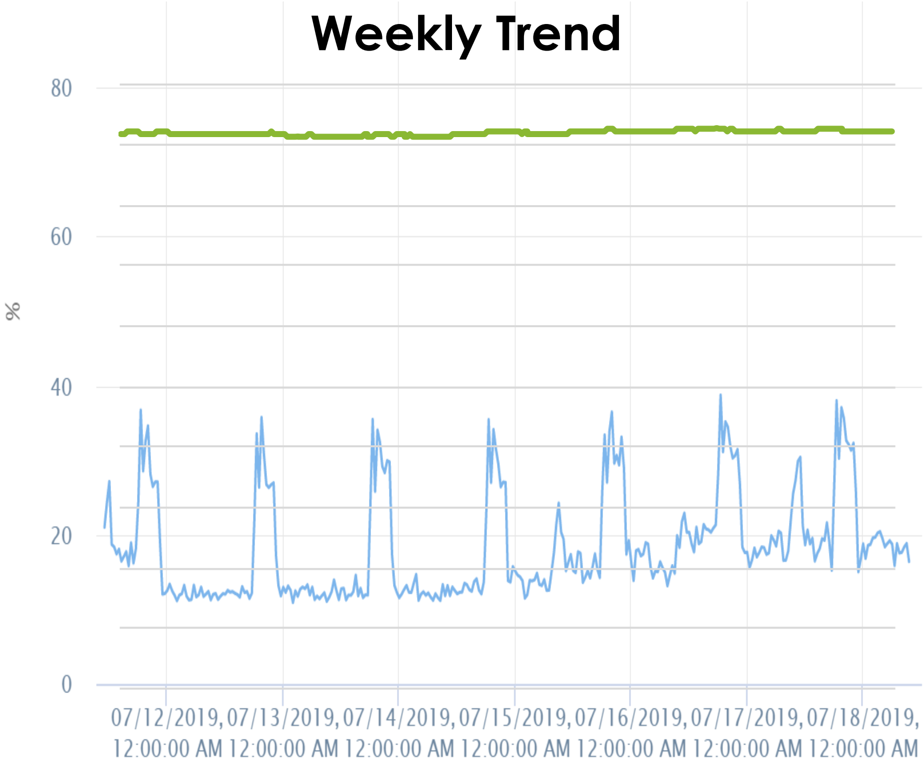
- ECS/S3 Object Storage 55% capacity of 2.2PB
- Growth ~57TB/month

- DBs 74% capacity of 343TB, expanding 100TB
- DB growth ~4TB/month

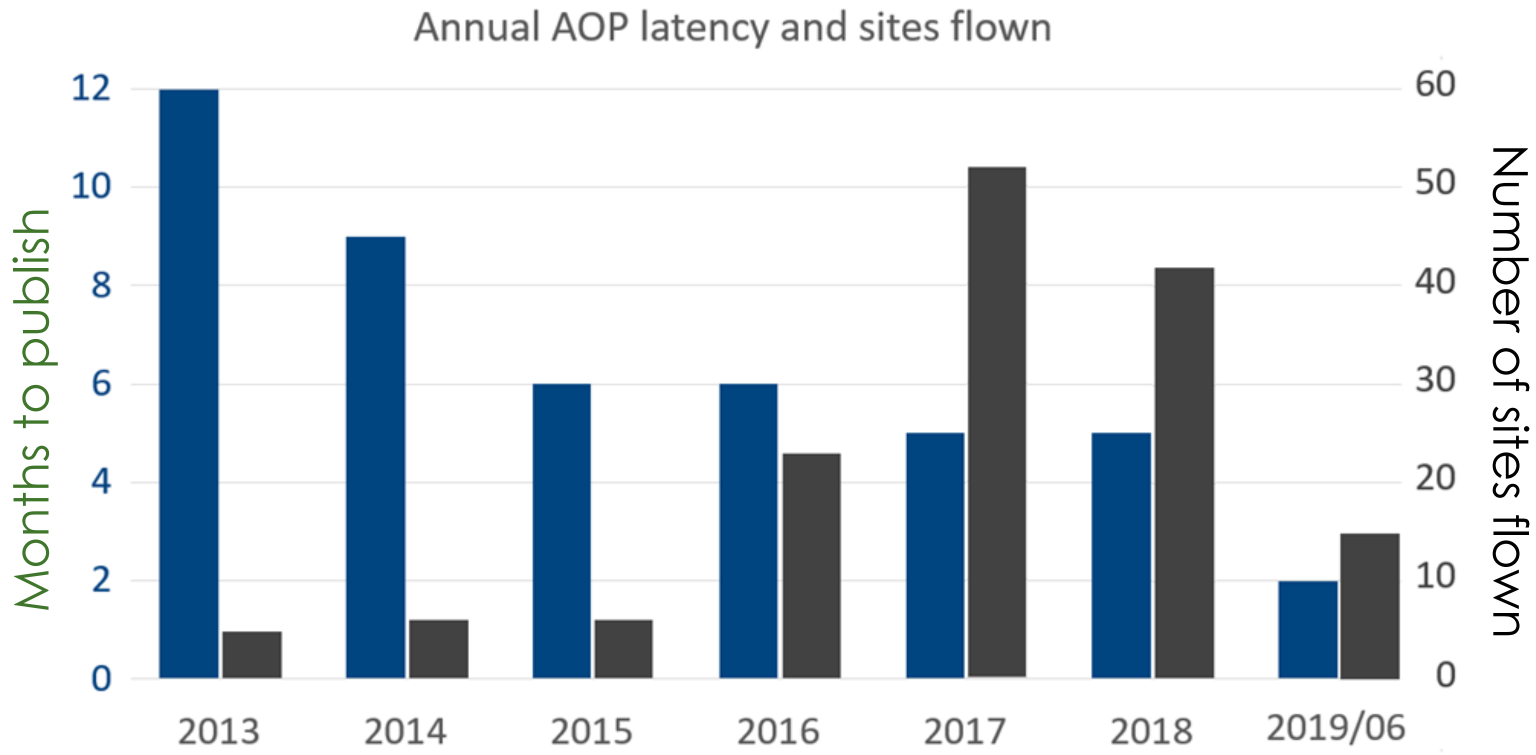


NEON CI Compute Capacity Utilization





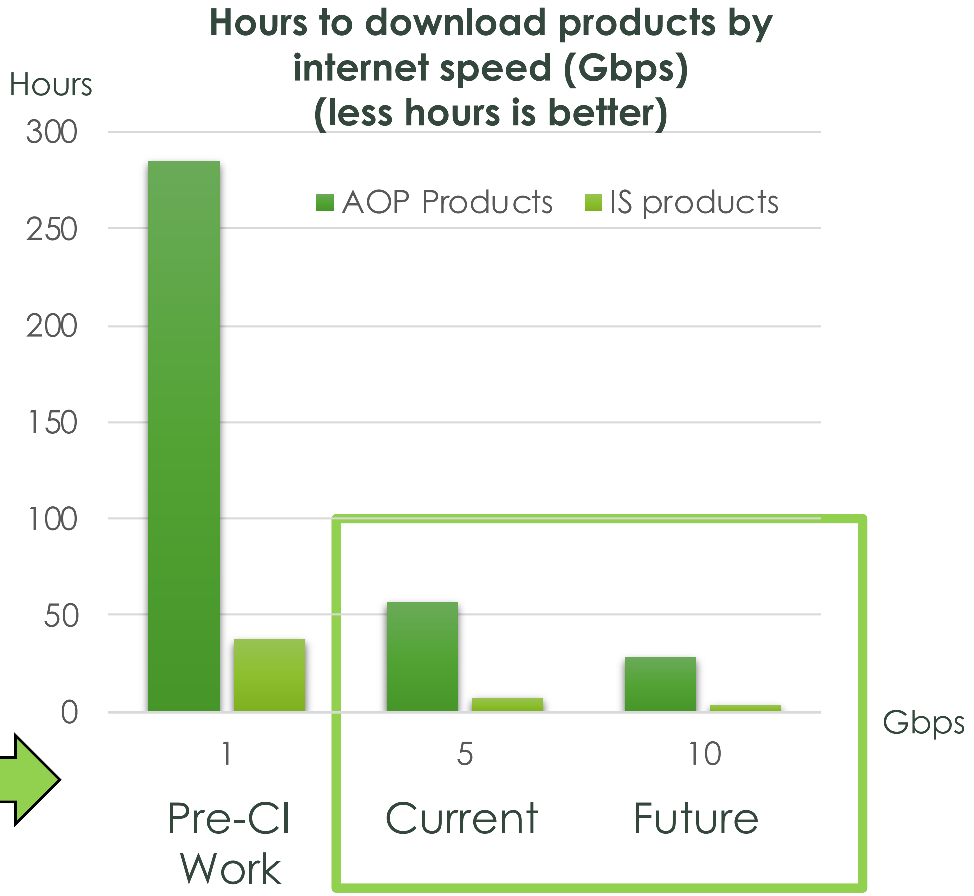
Aerial Platforms Data Processing Latency



NEON CI Connectivity Enhancements

- Firewalls upgraded from 1Gbps to 40Gbps
- Internet pipe upgraded from 1Gbps to 5Gbps

Time to download full AOP product **dropped from 12 days to 2.5 days**
Time to download IS product **dropped from 38 hours to 8 hours.**



NEON CI Messaging with Avro

neon
Operated by Battelle

NSF

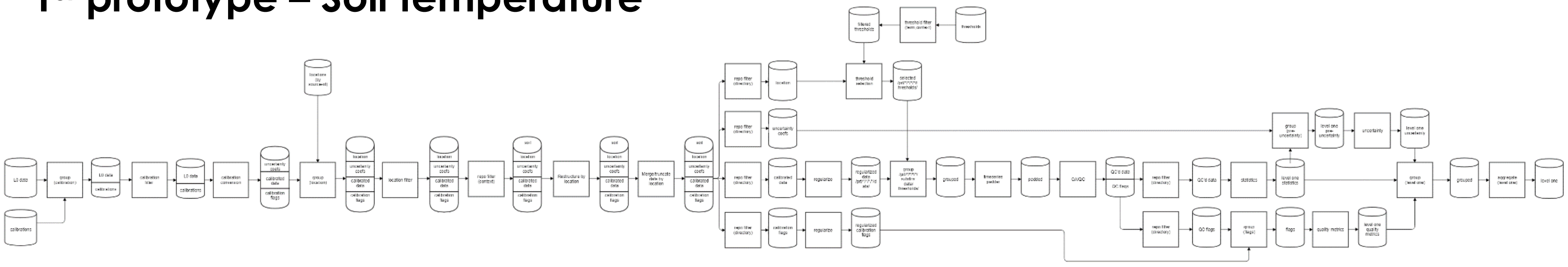
AVRO ™

- Standardized data serialization sy
  - Well documented, open source, maintained by Apache
  - APIs for many popular languages already exist
  - Already being used in “big data” platforms
- Rich file-based data storage structures
  - Fully self-describing data with no per-measurement overhead
  - Compact, fast, binary data format, with codecs for data compression
- Challenges
  - Sensor naming/model determined by manufacturer
  - Single schema to map part numbers to sensor types/assemblies
  - Interoperability of attribute nomenclature

NEON Sensor Processing Enhancement

NEED	SOLUTION
Automated response to data change (raw data, calibrations, location info, etc)	Pachyderm-based processing modules ‘listen’ for any data change
Traceability	Git-like version control for data and code
Reproducibility	Version-controlled Docker containers contain code and dependencies
Code re-usability	Highly modular processing design
Integrated Science-CI development	Docker-based, language-agnostic code packaging

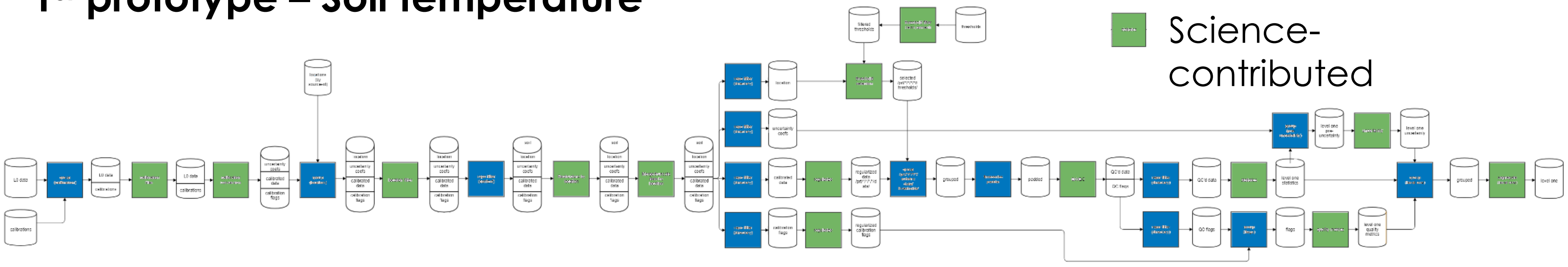
1st prototype – Soil temperature



NEON Sensor Processing Enhancement

NEED	SOLUTION
Automated response to data change (raw data, calibrations, location info, etc)	Pachyderm-based processing modules ‘listen’ for any data change
Traceability	Git-like version control for data and code
Reproducibility	Version-controlled Docker containers contain code and dependencies
Code re-usability	Highly modular processing design
Integrated Science-CI development	Docker-based, language-agnostic code packaging

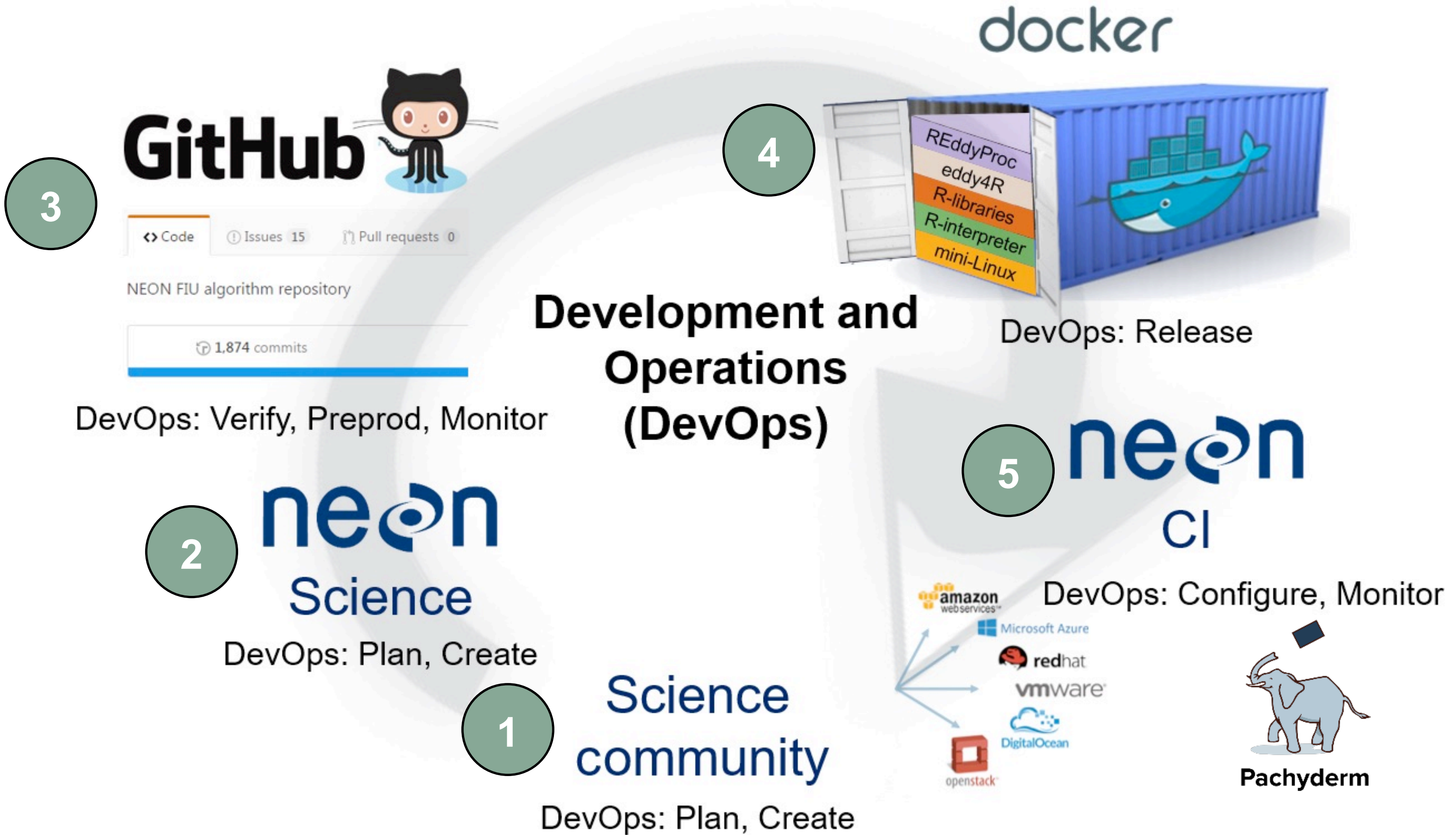
1st prototype – Soil temperature



🟦 CI-contributed

🟩 Science-contributed

Open-Source Data Pipelines



DevOps: Verify, Preprod, Monitor

DevOps: Release

Development and Operations (DevOps)

Science

DevOps: Plan, Create

CI

DevOps: Configure, Monitor

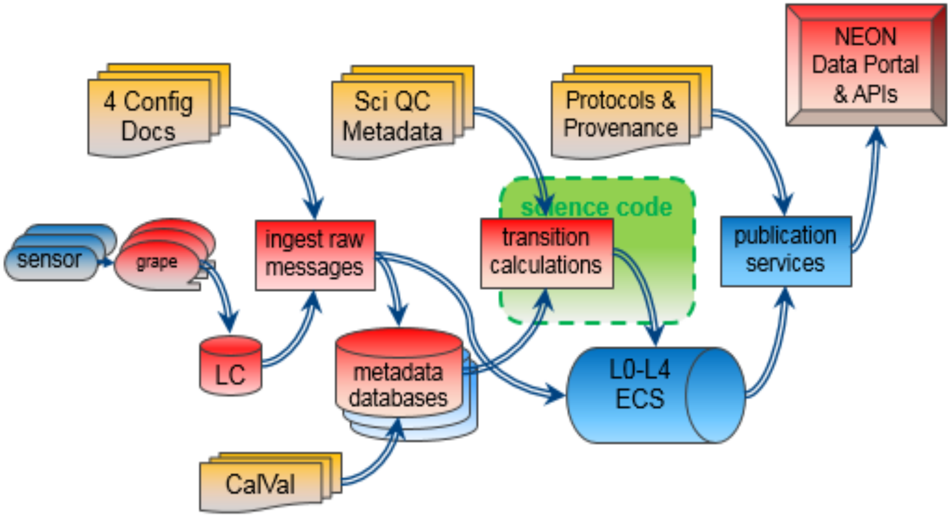
Science community

DevOps: Plan, Create

Pachyderm

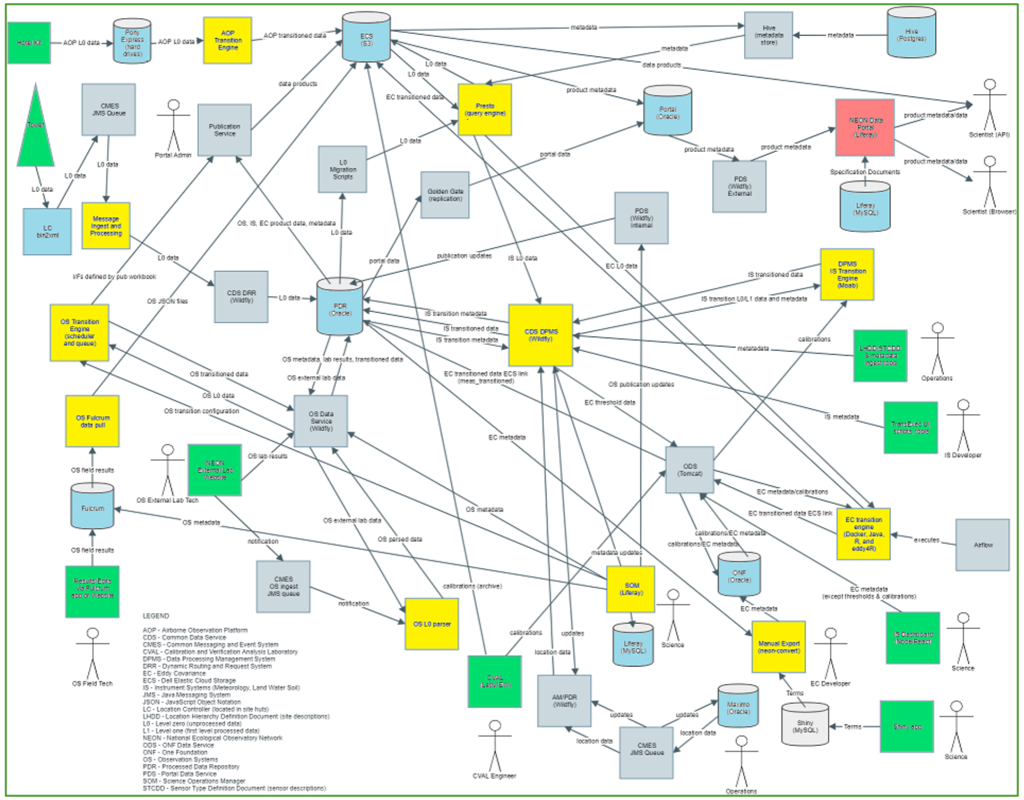
CoE Pilot Benefits to NEON Thus Far

- Short ramp-up due to receptivity/readiness to change
- Broadened network of expert CI colleagues
- Major upgrade to Data Portal's remote sensing visualization
- Accelerated Data Portal completion plan
- Affirmed strategies for workflow, messaging, & DR
- Raised critical mass of attention on semantics & schema.org
- Excited software developers
- Escalated accountability of CI
- More coming



Possible CoE Scope Amendments

- Methods for CI performance self-assessments
- Advice on CI documentation
- Consultation with CI development investors
- Inter-facility collaboration
- Workforce development?



LEGEND

AOP - Airborne Observation Platform
CDS - Common Data Service
CMES - Common Messaging and Event System
CVAL - Calibration and Verification Analysis Laboratory
DPMS - Data Processing Management System
DRR - Dynamic Routing and Request System
EC - Eddy Covariance
ECS - Dell Elastic Cloud Storage
IS - Instrument Systems (Meteorology, Land Water Soil)
JMS - Java Messaging System
JSON - JavaScript Object Notation
LC - Location Controller (located in site huts)
LHDD - Location Hierarchy Definition Document (site descriptions)
L0 - Level zero (unprocessed data)
L1 - Level one (first level processed data)
NEON - National Ecological Observatory Network
ODS - ONF Data Service
ONF - One Foundation
OS - Observation Systems
PDR - Processed Data Repository
PDS - Portal Data Service
SOM - Science Operations Manager
STCDD - Sensor Type Definition Document (sensor descriptions)

Engagement with Large Facilities
NEON Lessons Learned

1. ViSUS.org, CILogon, Schema.org, Baleni, DR template…

1. External dialog added valuable formality to planning
2. Project's readiness to improve accelerated idea exchange
3. Trust earned quickly
4. Schedule alignment near and long term nontrivial challenge
5. Awareness can always be broadened, & is worthwhile
6. Our proposed plans were fine, except those suboptimal ones
7. Funding horizons influence technical feasibility

Expanding Engagement with Large Facilities

- **Deep engagement**:
  - Identify a topic that is important and not-yet fully solved by the LF,
  - Conduct focused discussions, mix of virtual and in-person presence, hands-on work
  - Includes an engagement template that defines scope, sets expectations, identifies products
  - Work products: documents/papers, prototypes, schema implementations, demos
- **Topical discussions**:
  - Identify a topic that is important to a number of LFs
  - Facilitate virtual discussions, sessions at conferences, collect and share experiences, distill best practices
  - Discover opportunities for shared infrastructure
- **Community building:**
  - Identify related efforts
  - Collect information and disseminate information about the broad community activities
  - Maintain a living resource for community information
  - Develop new partnerships
- **Each engagement has a working group with 1-2 leaders and a set of work products.**

How to Engage?

We want to engage with you!

- http://cicoe-pilot.org
- **ci-coe-pilot@isi.edu**
- **Ewa Deelman deelman@isi.edu**
- **Participate in workshops and user surveys**

CI CoE PILOT

http://facilitiesci.org

Agenda

Monday

09:40 – 10:10	Guided Activity Kate Keahey and Rafael Ferreira da Silva
10:10 – 10:40	Break
10:40 – 12:00	Panel: State and Future of Cyberinfrastructure for Large Facilities Moderator: Dan Stanzione Panelists: Stuart Anderson (LIGO), Margaret Johnson (LSST) and Eric Lyons (Cyverse)
12:00 – 13:00	Lunch Break
13:00 – 13:15	NSF/CISE Perspective Erwin Gianchandani (NSF)
13:15 – 13:45	Large Facilities Data Lifecycle Anirban Mandal
13:45 – 15:25	Lightning Talks
15:15 - 15:30	Result Survey Overview and Setting up the Breakouts Ewa Deelman
15:30 – 16:00	Break
16:00 – 17:30	Parallel Breakouts: Collaboration, Technical and Non-technical CI challenges
17:30 – 18:00	Breakout Summaries Top 3-5 findings and recommendations from each group
18:30 – 20:30	Reception with cash bar

Tuesday

07:30 – 08:20	**Breakfast**
08:20 – 08:30	Setting the stage for Day 2 Ewa Deelman
08:30 – 10:00	Panel on Shared CI Services Opportunities and Challenges Moderator: Adam Bolton Panelists: Pamela Hill, JJ Kavelaars, Von Welch, and Mike Zentner
10:00 – 10:30	Break
10:30 – 12:00	Panel on Workforce Development and Retention Moderator: Tom Cheatham Panelists: Sharon Broude Geva, Frank Wuerthwein, Jim Rosser, Rachel Adams
12:00 – 13:00	Lunch Break
13:00 – 14:30	Parallel Breakouts: Community Building, Workforce, Ci landscape
14:30 – 15:00	Breakout Summaries Top 3-5 findings and recommendations from each group
15:00 – 15:15	Wrap-up

Notes

Please make comments / take notes during the workshop:

https://tinyurl.com/lf-ci-notes

Please start your text with **[your name]**

Don’t forget to check out the calling cards: main page
http://facilitiesci.org