2019 NSF Workshop on Connecting Large Facilities and Cyberinfrastructure

September 16, 2019 – AM Notes

Opening Remarks

1. NSF is interested in integrative, multi-disciplinary CI investments that help multiple facilities. Facilities without science commonalities can still have CI commonalities.
2. Workshop is meant to bring two communities together, foster exchange between LFs. Inform NSF on how to shape investments in CI. NSF is interested in integrative CI investments (simultaneously benefit multiple facilities). It is already happening, but want to scale it up.
3. Even if a physics, ecology facilities don't share science goals, they can share CI goals. Encourage identifying such solutions - NSF will be interested in hearing this.

Building LF Cyberinfrastructure Communities to Advance the Endless Frontier

1. Investing in big data at large scales: NEON and OOI examples. Centralized and 'distributed but integrated' observatory systems.
2. LF/CI challenges:
    a. Very large and diverse data sets; heterogeneous data on different temporal and spatial scales
    b. In-house software at separate facilities with very little sharing among them
    c. Need CI that is interoperable; need open communication and coordination and information/expertise sharing
    d. LFs may tend to reinvent the wheel (e.g., software) rather than building on existing successful work, largely because everyone is busy trying to keep up and so they end up working independently from each other
3. Starting conversations:
    a. Facility-led community engagement;
    b. Share through workshops and meetings;
    c. Bring together advisory group from different facilities
4. Connecting data streams (LIGO example)
    a. Facilities need to go beyond working in parallel
    b. Need to connect diverse investments
    c. Connect data streams from NSF and beyond

    d. Eg. NCAR and NEON are already communicating; geosciences and ecology; recent joint workshop; bringing addition power to NEON and NCAR data for each other;
5. The importance of finding solutions that improve the overall working of LFs, their CIs, and the achievement of their research missions. Example of being able to predict in order to do more useful work, e.g., bark beetle damage to forests
6. Encouraged us to proceed with the work of this workshop. The importance of this workshop as an early step toward overcoming many of these challenges. The importance of continuing to discuss the issues and identify ways to learn more and identify possible solutions.
    a. Developing LF CI standard principles
        i. 2017 LF CI workshop
        ii. 2018 - 2020 CICoE Pilot
7. Next challenges in BIO (Predictive is the keyword)
    a. Earth systems predictability
    b. Ecological forecasting

Questions/Comments

- Question: What does NSF plan to do to deal with the fact that each LF develops things differently? How do we really make this cohesive LF/CI happen well? It can't be all kumbaya, workshops, and carrots. Maybe we need sticks?
  Internal discussions within NSF are ongoing. They are cognizant of cultural differences across LFs. Mandates could be set: e.g., when setting up a new LF, requirements for a standard CI that the LF has to use. Reasonable standards that NSF wants to set. NSF will be informed about the standards and models through these workshops. However, first we need to gather more info and table enforcement for now. This workshop is intended to do that. Suggested discussion of models we need to incorporate in the facilitation, cooperative agreements, etc. She encouraged us to discuss the barriers around better collaboration and info sharing.
- Frank W suggested analyzing "structural self-interest." What are the things that work against us really collaborating and sharing info? That will then help us identify where we can put incentives. Suggested that incentives are more productive than sticks.
- Manish P: Pull from new use cases MM astrophysics. Application use cases is another incentive. Identify examples of success that can be emulated.
- Commenter emphasized the challenges around and the importance of data archives and their longevity. Joann talked about one of the central problems in archives—how to preserve the data in a way so it is still accessible and useable in the future.

Setting the Stage: 2017 CI workshop and the Cyberinfrastructure Center of Excellence Pilot

1. The workshop has both technical (e.g., challenges, LF CI and end user CI, better investments) as socio-technical topics (e.g., workforce, non-technical issues, etc.)
2. Previous workshop key findings:
    a. Close interactions, collaboration, and sharing
    b. Expertise, technical solutions, best practices and innovations
    c. Lack of easily accessible information about current CI technologies, solutions, practices, and experiences
    d. Lack of focused entity that could facilitate interactions
    e. Workforce development, training, retention, career paths, and diversity
3. CICoE Pilot: http://cicoe-pilot.org
    a. Build a blueprint for the CICoE with the LF and CI communities
4. Lessons learned with engagement with LF (NEON)
    a. F2F discussions and meeting builds trusts and relationships
    b. Benefits of formalizing the engagement: expectations, timelines, resources to use
    c. Importance of LF priorities and challenges, importance of good timing
    d. Organizing work around working groups and work products
    e. Be open to learn about what works, don'' fix it
5. NEON
    a. CI storage utilization: couple of TBs (Storage growth ~57TB/month, DB growth ~4TB/month)
    b. CI compute capacity utilization: CPU around 30-40% (annual), RAM 70-80% (annual)
    c. Organization willingness to change: big challenge
    d. Important to draw the line that the CoE is there to help and not evaluate
    e. Workforce: they cannot hire enough people to do the work, but can find them across other facilities
    f. Connectivity enhancement: down to few hours instead of ~270h
    g. CI Messaging with Avro (standardized data serialization system)
    h. Where the CI community stops and the user community starts?
    i. Sensor processing enhancement
        i. Need: automated response to data change; solution: pachyderm-based processing modules 'listen' for any data change
        ii. Need: traceability; solution: git-like version control for data and code
        iii. Need: reproducibility; solution: version-controlled Docker containers contain code and dependencies
6. CICoE benefits to NEON
    a. Short ramp-up due to receptivity/readiness to change

    b. Broadened network of expert CI colleagues
    c. Major upgrade to data portal's remote sensing visualization
    d. Accelerated data portal completion plan
    e. Affirmed strategies for workflow, messaging, & DR
    f. Raised critical mass of attention on semantics & schema.org
    g. Excited software developers
    h. Escalated accountability of CI
    i. M\ore coming
7. Possible CoE Scope Amendments
    a. Methods for CI performance self-assessments
    b. Advice on CI documentation
    c. Consultation with CI development investors
    d. Inter-facility collaboration (OOI Best practices paper: something to follow up)
    e. Workforce development? (needs additional work and collaboration with other facilities)

Qn: Few words about sustainability?
Qn: Given the turmoil in NEON over the firing of senior science staff, the resignation of the science lead, and the resignation of the scientific advisory board (https://www.sciencemag.org/news/2019/01/neon-ecological-laboratory-risk-fired-advisers-warn-nsf-after-shakeup), what lessons were learned? Specifically, how does a CI project tasked with supporting both the Large Facility and enabling the scientific community do big science navigate such challenges?

Sustainability is difficult; Alignment of funding; A center save funding by amplification; Multiple peoples' efforts; NEON's report about this will be very useful;
MP: In 2017 LF CI workshop; Pilot was a test run; If the Pilot is a good model, then how can we build it from there and make it more sustainable ?

Guided Activity

1. What are the most significant challenges faced by LFs or projects ?
    a. One set of suggestions:
        i. How do large facilities determine which pieces of their problem fall under "commodity CI" that can be shared versus the "speciality CI" specific to the science?
        ii. Internally, too much focus on "tools" to be used instead of "processes" that arrive at solutions.
        iii. It's difficult to identify long-term CI partnerships -- large facilities have very long lifespans compared to the existing CI providers (particularly, hardware investments). There is no long-term "fixed point" of shared CI providers.
2. What are the most important problems a CI CoE could solve?

    a. One set:
        i. Providing a translation layer to bridge the language/communication barrier between facilities and CI providers; creating a common taxonomy and vocabulary.
        ii. Helping to clarify the distinctions, possibly case-by-case, between common/commodity CI solutions versus aspects that are domain-specific, and clarifying the distinction between wants and needs of facilities.
3. Providing processes for engagement and integration between facilities and CI providers that accunt for operational stage and timing, and that are different depending on whether the need is for general consulting versus for specific solutions/services.
    a. Helping scientists identify and adopt new technologies (FPGA, GPUs) that could benefit/accelerate more effective research, rather than just trying to throw more of the old at it. More difficult for scientists to identify new technology solutions. We need to lower the barrier to adoption of tools and CI solutions.
    b. Trying to maintain a small staff with a large number of skills (especially specialized CI skills).
    c. Building lots of bespoke systems that outlive personnel turnover, how to manage these in a way that causes less loss of maintainability.
    d. Navigating the line between PIs not wanting to share data and eg NSF requirement on LF that users publish. There are 2 challenge avenues to this:
        i. Cultural/generational change
        ii. Pragmatic challenge: ~"I could publish faster if I had more resources"
    e. Challenges faced by large facilities:
        i. Balance between sustaining what has been created vs engagement in new activities (e.g., collaborations and synergies)
            1. Eg. Revamping existing infrastructure is a drain on existing resources for LFs/academic CI projects;
            2. The tension between being funded to do something new vs. sustaining existing infrastructure.
        ii. Time to revamp and update existing infrastructure
            1. Updates in industry increase revenue (academics have little incentives)
        iii. Many NSF projects are funded to do "new" science/development, not update the old
    f. CyberSecurity, Standard Operating Procedures, DataStorage/transfer over the a diverse, dynamic, LF (ARF).
    g. Biggest challenge for NHERI is staffing: need matched matrix of "distribution of CI skill requirements vs. distribution of staff"; Another example is "No local CI staff"
    h. Automation of knowledge transfer for long projects like in the LFs. There is turnover during the project period and knowledge is lost when a particular project member moves on. Configuration management and automation are desirable.

4. What are the most important problems a cyberinfrastructure center of excellence could solve?
    a. Regular Assessment of All/Many Large Facilities to identify common needs and problems
    b. Develop a suite of turn-key solutions, recommendations, for the annual problem
    c. Training: Training is focused on those emerging solutions (as opposed to best practices which can be captured)
    d. Discussion/tension among the following structures:
        i. CoE provides training and recommendations, facilities stands up the hardware and software
        ii. CoE provides the software, services, and hardware, paid for by NSF and facilities uses it
        iii. CoE provides the software, services, and hardware, paid for by the facilities (and written into their proposal)
    e. Maintain expertise in specialized areas where individual facilities cannot afford to do so due to needs being episodic rather than constant.
    f. Vet new technologies and help LFs understand what is applicable to their use cases.
    g. CoE could be an incentive for collaboration by providing an ‘army’ of skills
        i. CoE uses a competitive process to select which facility/ies (N facilities coming together) ‘wins’ the benefit of CoE inputting effort to complete project X with efforts/deliverables/timelines; Augmented by effort and matching funding from both sides
    h. Provision of Standards
        i. not auditing //Eg bad example: NSF branding requirement was done badly - hard line wasn't a good approach
    i. Knowledge base
    j. Skills base that can be hired as a consultation. Should be free for LFs
        i. And via that process skills transfer to the LF
    k. Provide training that a specific LF doesn't have time. Be a center of excellence for transferring skills to
        i. Eg: Embedded systems
        ii. Eg: Trusted CI workshop breakout eg had a session on embedded security
        iii. Monthly seminars, online webinars
    l. Work on what barriers exist for LFs to use new tech
    m. Concern: CoE can be in competition with LFs in some aspects; Need separating operation of LFs and development of software by CoE
    n. CoE also needs to fit in with the culture of the LF and align with timing of funding horizons
    o. CoE can be “Home for first volunteer catalog of existing and upcoming CI solutions”
        i. Software, consulting; eg. upgrade consulting

p. CoE can evaluate NSF mandated proposal requirements with respect to sharing the CI for LFs

Panel: State and Future of Cyberinfrastructure for Large Facilities

1. What has changed since 2017?
    a. Dealing with interactive computing…batch tools vs VMs, vs Shiny/Jupyter/etc
    b. Difficulties in changing…
        i. facilities take on building lots of own bespoke components.
        ii. Lots of "I have a hammer, so this looks like a nail" → Projects become limited by that frame of reference, may lose sight of the business purpose.
    c. More effectively getting projects to contribute to or use shared systems.
    d. Growing emphasis on cyberinfrastructure due to time-to-solution concerns.
    e. More distributed computing, less in-house, going into production
2. What are the major challenges today and 5 years from now?
    a. How do we update CI in production without disrupting the science mission?
    b. Workforce sustainability
    c. Training, scaling workforce, preparing for new technologies coming down the pipe
        i. How do we leverage one another's strengths within the community to cross-train, help one another keep up
        ii. "long tail" of people who want to start catching up keeps getting longer as technology moves faster
    d. Difficult to make changes to MREFCs in progress, even when the technology landscape is changing under you.
        i. during conceptualization phase, it's especially important to bring in CIexpertise, people with a frame of reference for operations, reuse of CI capabilities
        ii. during operations, chance to re-integrate with broader CI community… look for how to evolve capabilities, adopt best practices, contribute to shared capabilities, expertise practices
3. What can be done at no additional cost?
    a. ensure healthy balance of staffing, especially from conceptualization phase
    b. set up distinct core operations earlier in MREFC to ensure mission focus, incrementally transition to operations
    c. Training and workshops
    d. Scaling people
    e. leverage one another's storytelling/marketing engines
    f. Leverage community training materials: Carpentries and DataCamp etc - Utilise these to provide/share/give credit
4. Where can collaborations and sharing help? (hw, sw, people)

5. Compound interest on technical debt is EXPENSIVE!
6. What are the current blockers to collaboration?
    a. cost of innovation is now bigger than “not invented here”...so busy keeping things running that it’s hard to invest in new collaborations, have to be choosy
    b. Balance early on is needed...there has to be some scope for innovation, get people out of silos? ← I didn’t entirely grasp where the speaker was going here
    c. It’s too easy to miss the hidden costs...small changes may be more labor-intensive than expected (sounds like a lack of code compartmentation, architecture issues, lack of comprehensive testing that makes changes like new IdM systems hard) Technical debt again…
    d. A more agile construction phase would be more efficient, but how to get there? Right now, adopting common tools is MUCH easier in the ops phase than in construction…
        i. software built in construction of LIGO was thrown away, didn’t know enough yet at conceptualization, stuck with it in construction, fixed it in ops
7. CI fundamental principles behind constructing the facilities vs. buzzwords and technologies
8. What are the things that remain the same as we evolve ?
9. Cyverse, as we moved forward, because of funding ramp-down, how do we decide what things fall off;
10. Need taxonomy of logical architectures, not specifics
11. Quantifying the cost of changes is important;
12. Entering an era of we don’t have enough computing; Data coming in outweighs computing capacity; How flexible is the thinking in the way computing is done - can we do things in 16 bit ?
13. “Time to science”;
14. Challenges that heterogeneity implies...trying to build abstractions around things that we have...how much pain does platform heterogeneity inject as it grows, do we adapt or throw up our hands and reject it?
    a. Some low-resource projects can keep running on legacy systems.
    b. Some groups are actively moving to new architectures.
15. Problems/challenges: Software vs. data vs. workforce
    a. Cyverse: 10% data; software and people rest half
    b. LIGO: 70% workforce, 20% software; 10% data
    c. LSST: Maximum goes to data;