Northeast Cyberteam – Building an Environment for Sharing Best Practices and Solutions for Research Computing

John Goodhue
Massachusetts Green High Performance Computing Center
Holyoke, MA, USA
jtgoodhue@mghpcc.org

Julie Ma
Massachusetts Green High Performance Computing Center
Holyoke, MA, USA
jma@mghpcc.org

Adrian Del Masetro
University of Vermont
Burlington, VT, USA
adrian.delmaestro@uvm.edu

Sia Najafi
Worcester Polytechnic Institute
Worcester, MA, USA
snajafi@wpi.edu

Bruce Segee
University of Maine
Orono, ME, USA
segee@maine.edu

Scott Valcourt
University of New Hampshire
Durham, NH, USA
sav@unh.edu

Ralph Zottola
University of Alabama
Burmingham, AL, USA
rzottola@uab.edu

***Abstract*—The Northeast Cyberteam Program is a collaborative effort across Maine, New Hampshire, Vermont, and Massachusetts that seeks to assist researchers at small and medium-sized institutions in the region with making use of cyberinfrastructure, while simultaneously building the next generation of research computing facilitators. Recognizing that research computing facilitators are frequently in short supply, the program also places intentional emphasis on capturing and disseminating best practices in an effort to enable opportunities to leverage and build on existing solutions whenever practical. The program combines direct assistance to computationally intensive research projects; experiential learning opportunities that pair experienced mentors with students interested in research computing facilitation; sharing of resources and knowledge across large and small institutions; and tools that enable efficient oversight and possible replication of these ideas in other regions.**

Each project involves a researcher seeking to better utilize cyberinfrastructure in research, a student facilitator, and a mentor with relevant domain expertise. These individuals may be at the same institution or at separate institutions. The student works with the researcher and the mentor to become a bridge between the infrastructure and the research domain. Through this model, students receive training and opportunities that otherwise would not be available, research projects get taken to a higher level, and the effectiveness of the mentor is multiplied.

Providing tools to enable self-service learning is a key concept in our strategy to develop facilitators through experiential learning, recognizing that one of the most fundamental skills of successful facilitators is their ability to quickly learn enough about new domains and applications to be able draw parallels with their existing knowledge and help to solve the problem at hand. The Cyberteam Portal is used to access the self-service learning resources developed to provide just-in-time information delivery to participants as they embark on projects in unfamiliar domains, and also serves as a receptacle for best practices, tools, and techniques developed during a project. Tools include Ask.CI, an interactive site for questions and answers; a learning resources repository used to collect online training modules vetted by Cyberteam projects that provide starting points for subsequent projects or independent activities; and a Github repository. The Northeast Cyberteam was created with funding from the National Science Foundation, but has developed strategies for sustainable operations.

Keywords—workforce development, research computing facilitator, project portal, Ask.CI, MGHPCC, Northeast Cyberteam

I. Introduction

The Northeast Cyberteam [8] was established in 2017 through the development and funding of a National Science Foundation proposal led by the Massachusetts Green High Performance Computing Center (MGHPCC) with collaborators in Massachusetts, Vermont, New Hampshire, and Maine. The collaborators recognized a diversity of specialized knowledge at various institutions in the region as well as an abundance of researchers whose work would benefit by an increased use of large-scale cyberinfrastructure, but for whom making the transition required overcoming barriers such as not knowing how to start, who to talk to, or how to address problems that arose. The model of the Northeast Cyberteam is to identify research computing facilitators across the region who are willing to act as mentors for students that want to learn more about research computing. A typical project, proposed by a researcher, stems from the researcher's computational or data needs outstripping the processing power of a laptop, or the desire to apply a new technique (e.g., machine learning) in the researcher's domain. Each project involves a researcher seeking to better utilize cyberinfrastructure in research, a student facilitator, and a mentor with relevant domain expertise. These individuals may be at the same institution or at separate institutions. The student works with the researcher and the mentor to become a bridge between the infrastructure and the research domain. Throughout the project, emphasis is placed on documenting the project, tools used, and lessons learned.

This work was partially support by the National Science Foundation under grant award ACI-1659377.

978-1-7281-9219-2/20/$31.00 ©2020 IEEE

Through this model, students receive training and opportunities that otherwise would not be available in both the research domain and the cyberinfrastructure domain, research projects get taken to a higher level, the effectiveness of the mentor is multiplied, and value is created for the researcher AND the research computing community.

II. Program Motivation

The Northeast Cyberteam model was developed for several distinct and related reasons. The program creates opportunities for students to learn about research computing facilitation and for researchers to learn how to make use of cyberinfrastructure that is available throughout the region. The program also intentionally places emphasis on capturing and disseminating best practices.

A. *Creating Opportunities for Students*

Research computing facilitators have colloquially been referred to as "purple unicorns" based on their rarity. Researchers with domain expertise exist in abundance, but those researchers can and should focus on their research areas as opposed to the tools that they use. Similarly, the staff that support the computational infrastructure across all domains rarely have the domain specific expertise (or time) to provide specialized support. Student facilitators have the energy and motivation to successfully fill this gap. Through a combination of real-world problems and interactions with mentors, these students gain research computing knowledge relevant to the domain, and domain knowledge relevant to the research computing tools and techniques being used. Northeast Cyberteam projects are typically a few months in duration and have specific endpoints that are included as part of the proposal process. This is in contrast to conventional open-ended research projects, and has the desirable side benefit of identifying a publishable result for the student.

B. *Creating Opportunities for Researchers to Publish the Research Computing Angle of Their Work*

A conventional scientific workflow for a domain specific project may focus entirely on the final outcome and results and thus may not sufficiently record or optimize the computational framework that produced it. This is a result of a myriad of factors including publication pressure; lack of formal computational training; use of scientific legacy codes; absence of a clear incentive structure; lack of knowledge of research computing as a separate discipline; and others. This ad hoc approach to research computing across scientific disciplines can lead to the generation of errors; a reduction of reproducibility and associated data provenance; lost opportunities to leverage previous work; and ultimately hinder scientific progress. One of the goals of the Northeast Cyberteam Initiative is to address these underlying causes, while remaining mindful of impediments to the adoption of research computing best practices.

C. *Capturing/Communicating Best Practices to Prevent "Reinventing the Wheel"*

As noted above, one of the key focus areas of the Northeast Cyberteam has been to capture practices and protocols, developed as a part of the projects, that can be used as a starting point for future work with similar parameters. Given the aforementioned scarcity of experienced research computing facilitators, and the ever-expanding range of opportunities to apply research computing, leveraging previously developed research computing techniques is an effective strategy to advance science in a timely manner while moderating facilitator workload.

D. *Computational Physics - A Case Study*

Computational physics is now well-established as both a separate research discipline, and also as an essential bridge between experimental and theoretical research. The community has deep and significant knowledge of software engineering and is behind many large-scale open source projects that support 1000s of users including ALPS -- the Algorithmic Library for Physics Simulations [1], electronic structure codes such as Quantum Espresso [2] and the molecular dynamics software GROMACS [3] to name just a few in the space of quantum condensed matter and materials physics. However, as computational tools become integrated across all areas of physics, especially approaches such as deep learning, there is a real need for a low-barrier strategy to communicate best-practices and train the actual practitioners, who are often undergraduate and graduate student researchers. Currently this training is not performed in the classroom, and is a combination of informal localized research group discussions (i.e. between senior and junior members), web searches, and pages such as stackoverflow. While this is certainly useful, the resulting practices and outcomes are not widely communicated and thus successful strategies are not often formally disseminated or reproduced. There is also a considerable risk of knowledge loss when local experts graduate or move on from a research group without leaving sufficient documentation for the next generation.

We have observed that participation in the Cyberteam for research computing facilitators provides a forum to share this information as well as learn from a broader community of practitioners. Moreover, students have been provided opportunities to present their work at conferences specializing in research computing (e.g. PEARC, SC, HPEC) offering significant bi-directional benefits including enhanced knowledge, but also improved and broadened dissemination of scientific results.

Similar examples can be found in fields such as bioinformatics and computational chemistry. Our most recent projects have even taken this notion to an ambitious and sophisticated art project that wishes to take advantage of tools and techniques that have been developed to leverage advances in genetic algorithms and machine learning. This project intends to generate procedural artistic works, but has the more ambitious goal to investigate and interrogate how deep neural network hidden layers may discover new forms of artistic expression. This is in analogy to AlphaZero learning and employing previously undiscovered strategies in the game of Go [4].

III. Tools/Platform

As noted above, building self-service learning skills is key to our strategy to develop facilitators through experiential learning, recognizing that one of the most fundamental

978-1-7281-9219-2/20/$31.00 ©2020 IEEE

characteristics of successful facilitators is their ability to quickly learn enough about new domains and applications to then be able draw parallels with their existing knowledge and help to solve the problem at hand. The program also intentionally places emphasis on capturing and disseminating best practices. The platform and tools we have developed to support this approach are described below.

A. Cyberteam Portal

The Cyberteam Portal (http://necyberteam.org) is used to access self-service learning resources that provide just in time information delivery to participants as they embark on projects in unfamiliar domains. The goal of these learning resources is to reduce the need for direct assistance; reduce duplication of effort by adapting and building awareness of available documentation, training, application software and software utilities; and supplement these resources where there are high impact opportunities.

B. Ask.CI

One of the first tools that we developed under the auspices of the Northeast Cyberteam Program was Ask.cyberinfrastructure.org (Ask.CI), a collaborative, crowd-sourced Q&A site curated specifically for the research computing community. [7] The project began in September 2017 with the vision of constructing a resource that allows the research computing community to achieve better and faster science results by making answers to commonly asked questions readily accessible from a shared, public knowledge base. While it started as an effort to support the Northeast Cyberteam participants, it quickly became clear that it was a tool that would be beneficial to and would derive benefit from the research computing community at large.

Establishing a Q&A site of this nature requires some tenacity. We have gained traction and hope to continue to engage the broader community to firmly establish this platform as a tool for the global research computing community. The goal of the project is to aggregate answers to a broad spectrum of questions that are commonly asked as researchers and educators utilize advanced computing and data resources, creating a self-service knowledge base for the domain researchers, facilitators, cyberinfrastructure engineers and others that comprise the research computing community. The hope is that not only will Ask.CI become a great resource for the community, but that it will also provide public testimony of the importance of research computing and how it exists in relation to enterprise IT, computer science, and domain research.

C. Learning Resources Repository

Early on in the project, we also identified the need to curate and reference the best of the research computing training resources available on the web. This relatively static information, such as introductory training modules on Linux clusters, programming languages, and schedulers, were useful knowledge pointers that did not require constant updating. Yet, many of these resources were difficult to find using common search engines, and those search engines often returned results that were not associated with the problem that the technology community was attempting to solve. The resource repository that we developed is designed to help facilitators come up to speed on particular topics when needed by providing pointers to publicly available, relevant, and vetted training resources. The modules that we are collecting are self-paced, and clearly defined, requiring varying levels of expertise. As a result, key topical information can be readily discovered through tags in the Learning Resources Repository and facilitators know that others in their position have already considered these training resources to be the most valuable.

D. Launch and Wrap Presentations

At the beginning of every Cyberteam project, we ask the students to present a "Launch" presentation which includes a brief introduction to their project and goals for both the project and for themselves, at a monthly meeting of all active student/mentor/researcher teams associated with the Northeast Cyberteam. At the conclusion of the project, we ask the students and mentors to summarize their findings in a "Wrap" presentation. Through this process, and using standard templates, every project is asked to identify what new learning took place, where is that learning curated, any publications of code generated, and anything that contributes to the scientific domain of the project. Students who engage in this exercise find that their generated project output, when summarized in this fashion, is one step from converting to a formal publishable format such as a poster or paper for a professional conference. This exercise encourages students to become academic publishers and contribute to the body of knowledge available to the research community.

E. Github Repository

Ensuring that the code artifacts developed by Cyberteam projects are not lost to one-time project uses, the Cyberteam program has developed a Github repository into which any code base that has been developed has a formal, available home for potential reuse by other researchers and facilitators. Typically, containerized codes and scripts have been deposited from more recent projects, but storing any type of code that could be reused to address a similar problem in the community is encouraged.

IV. Example Publications

The Northeast Cyberteam has created opportunities for student facilitators and their researchers to publish results that highlight the research computing aspect of their work as well as the tools and methods developed in support of the process. A few recent examples are described below.

A. PEARC Posters

Since inception, the Northeast Cyberteam has encouraged researchers, mentors, and especially students to communicate the work they have accomplished within the Cyberteam framework. Each year, at least one student poster was accepted at the PEARC conference, and the Cyberteam program has been able to provide travel and conference fee support to ensure that the students are able to attend in person and learn about the research computing practice. For PEARC18, the first PEARC that took place after the Northeast Cyberteam program commenced, we submitted a student poster called "Adventures of Two Student Research Computing Facilitators" that described the Northeast Cyberteam process and two Cyberteam projects [9]. The poster was accepted with high marks and reviewers noted that each project had enough scientific merit and significance to justify its own submission. Based on that feedback, in subsequent years, we published posters about individual student projects. [5, 10, 11] In 2020, for example, a student project from the University of New Hampshire was accepted for the PEARC20 conference titled "Augmented Reality: Telehealth Demonstration Application," that details the development process involved in creating an augmented reality (AR) proof-of-concept application to include a care plan tracker established by a patient's doctor to allow the patient to do daily tasks without a health care worker's supervision using a Microsoft HoloLens device. [5] Reviewers rated this poster with the highest marks, specifically commenting that projects such as these are rarely presented at this conference, but should be highlighted to serve as examples of the work that advanced researching computing supports.

B. *Gulf of Maine 2050 Posters*

The "Tracing Oceanic Pathways" project explored the productive waters of the Gulf of Maine (GOM) which support one of the most biodiverse and economically important marine habitats in the world. This productivity relies on adequate concentrations of nutrients being transported into the photic zone during times of the year when significant amounts of sunlight are present. As such, an understanding of the nature and variability of the pathways that deliver nutrients to the photic zone is desirable. In this project, legacy FORTRAN codes which model these trajectories were ported to a linux environment and containerized to run easily on a cluster operated by the Massachusetts Institute of Technology. This allowed the researcher to obtain trajectory data that had previously been unavailable due to insufficient compute and storage resources. It also allowed the team to form relationships with potential collaborators based on their ability to run trajectories on the cluster. The student facilitator, a computational chemistry major, learned Fortran, containers and oceanographics. The researcher, Dr. Kristine Burkholder and the students in her lab presented two posters at the Gulf of Maine 2050 conference based on the modeling capabilities developed in this project. [19, 20]

C. *PEARC/SC Birds of a Feather Sessions*

As described previously, Ask.CI, the Q&A site for the research computing community, is a project that grew out of the Northeast Cyberteam but rapidly became a resource for the research computing community at large. We have presented Birds of a Feather (BOF) sessions at every PEARC and Supercomputing conference since launching Ask.CI at PEARC18 [12, 13, 14, 15, 16], to educate the community about Ask.CI and to encourage participation. Beginning in 2018, we also presented BOFs about the Northeast Cyberteam and the Cyberteam portal [17, 18] in a successful effort to find collaborators to make use of the portal in other Cyberteam efforts across the country, and to replicate and expand on the process.

D. *PEARC Papers*

The PEARC20 paper titled, "Northeast Cyberteam:Workforce Development for Research Computing at Small and Mid-sized Institutions," details the Cyberteam project's purpose in meeting the need for support staff who can help faculty make the best use of available computing resources at small and mid-sized institutions, which is a significant challenge. [6] By building a pool of research computing facilitators that can be shared across institutional boundaries while also developing self-service tools that reduce the support burden, the Cyberteam has been able to make a change that is being felt across the country. This paper built on work published at the SEHET workshop co-located at PEARC19 the previous year.

V. Conclusions

While the Northeast Cyberteam program approaches its performance conclusion window, the outcomes and artifacts generated continue to have a promising future. Other Cyberteam project awardees from the National Science Foundation have already declared their intention to adopt the tools and methodologies of the Northeast Cyberteam and have begun the integration process. In doing so, these additional programs in new states and regions of the country will generate new content on Ask.CI, and new training modules in the Learning Resources Repository; launch new projects; involve more mentors and researchers, and create the opportunity for additional publishable materials in the research computing community. We look forward to new investigators following our methods to achieve similar outcomes and highlight our reproducible approach.

Acknowledgment

This work was partially supported by the National Science Foundation under grant award ACI-1659377. The authors thank the many Cyberteam mentors and student facilitators for their participation in this effort.

References

[1] Bauer, B, et al. "The ALPS Project Release 2.0: Open Source Software for Strongly Correlated Systems." *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2011, no. 05, 2011, doi:10.1088/1742-5468/2011/05/p05001.

[2] Giannozzi, P, et al. "Advanced Capabilities for Materials Modelling with Quantum ESPRESSO." *Journal of Physics: Condensed Matter*, vol. 29, no. 46, 2017, p. 465901., doi:10.1088/1361-648x/aa8f79.

[3] Abraham, Mark James, et al. "GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers." *SoftwareX*, vol. 1-2, 2015, pp. 19–25., doi:10.1016/j.softx.2015.06.001.

[4] D. Silver et al., "Mastering the game of Go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, 2017, doi: 10.1038/nature24270.

[5] Boyd, N., Hawkins, J., Yang, T., Valcourt, S. "Augmented Reality: Telehealth Demonstration Application." *Proceedings of the Conference on Practice and Experience in Advanced Research Computing (PEARC20)*, Portland, OR, July 26-30, 2020, doi:10.1145/3311790.3399629.

[6] Goodhue, J., Ma, J., Del Maestro, A., Najafi, S., Segee, B., Valcourt, S., Zottola, R. "Northeast Cyberteam:Workforce Development for Research Computing at Small and Mid-sized Institutions." *Proceedings of the Conference on Practice and Experience in Advanced Research Computing (PEARC20)*, Portland, OR, July 26-30, 2020, doi:10.1145/3311790.3396662.

[7] Ma, J., Culich, A., Sochat, V., Najafi, S., Segee, B., Braiterman, Z., Battelle, T., Goodhue, J., Brunson, D., Hill, C., Zottola, R., Singh, R., Bulekova, K., Pessin, J., Cheatham, T., DelMaestro, A., Valcourt, S., Thoelen, R., Smith, J. "Ask.Cyberinfrastructure.org: Creating a Platform for Self-Service Learning and Collaboration in the Rapidly Changing

Environment of Research Computing." SEHET20, Portland, OR, July 26-30, 2020. Unpublished.

[8] Goodhue, J., Ma, J., Del Maestro, A., Najafi, S., Segee, B., Valcourt, S., Zottola, R. "Northeast Cyberteam Program – A Workforce Development Strategy for Research Computing." *The Journal of Computational Science Education*, vol. 11, no. 1, 2020, pp. 8–11., doi:10.22369/issn.2153-4136/11/1/2.

[9] Howard, N., Colella, N (2018). "Adventures of Two Student Research Computing Facilitators." *Proceedings of the Conference on Practice and Experience in Advanced Research Computing (PEARC18)*. Pittsburgh, PA, July 22-26, 2018.

[10] Ben Burnett, Abigail Waters, Ben Levy, Julie Ma (2019). Creating a Cluster Oriented Workflow for Users of NetLogo. PEARC19. Chicago, IL.

[11] Michael Butler, Bruce Segee, and Liping Yu (2019). Facilitating high performance computing for the sciences: a case study in machine learning for materials science. PEARC19. Chicago, IL.

[12] Ma, J., Culich, A., Battelle, T., Brunson, D, Cheatham, T., Goodhue, J., Hill, C., Oleinik, K., Pessin, J., Smith, J., Yockel, S.. (2018). "On Launching a Research Compu8ng Stack Exchange or Similar Q&A Site" *Proceedings of the Conference on Practice and Experience in Advanced Research Computing (PEARC18)*. Pittsburgh, PA, July 22-26, 2018.

[13] Ma, J., Culich, A., Battelle, T., Brunson, D, Cheatham, T., Goodhue, J., Hill, C., Oleinik, K., Pessin, J., Smith, J., Yockel, S.. (2018). "On Launching Ask.CI, a Q&A Platform for Research Computing, Using StackExchange and Discourse." SC18. Dallas, TX. November 12-15, 2018.

[14] Ma, J., Battelle, T., Braiterman, Z., Brunson, D., Cheatham III, T., Culich, A., Goodhue, J., Hill, C., Oleinik, K., Pessin, Singh, R. J., Smith, J., Sochat, V, & Yockel, S. (2019). "Ask.CI, the Q&A site for Research Computing - Year 1 Lessons Learned." *Proceedings of the Conference on Practice and Experience in Advanced Research Computing (PEARC19)*. Chicago, IL.

[15] Ma, J., Culich, A., Battelle, T., Goodhue, J., Oleinik, K., Pessin, J., Sochat, V., Brunson, D., Hill, C., Cheatham, T., Braiterman, Z. (2019). "Ask.CI, the Q&A Platform for Research Computing - Recent Developments." SC19. Denver, CO. November 17-21, 2019.

[16] Ma, J., et al. "Ask.CI, the Q&A site for Research Computing - Year 2 Lessons Learned, Plans for Year 3." BOF Presentation, *Proceedings of the Conference on Practice and Experience in Advanced Research Computing (PEARC20)*, Portland, OR, July 26-30, 2020.

[17] Goodhue, J., Ma, J., DelMaestro, A., Najafi, S., Segee, B., Valcourt, S., Zottola, R. (2019). "Northeast Cyberteam Initiative - a Workforce Development Strategy for Research Computing at Small and Medium Sized Institutions". BOF Presentation, *Proceedings of the Practice and Experience in Advanced Research Computing (PEARC19)*. Chicago, IL.

[18] Ma, J., Goodhue, J., Chakravorty, D, DelMaestro, A., Griffoen, J., Knuth, S., Middelkoop, T., Najafi, S., Segee, B., Valcourt, S., Zottola, R. "The Cyberteam Portal, a Shared yet Independent Platform for Cyberteam Development." BOF Presentation, *Proceedings of the Conference on Practice and Experience in Advanced Research Computing (PEARC20)*, Portland, OR, July 26-30, 2020.

[19] Ladue, T., He, R., Burkholder, K. (2019). "Comparing Subsurface Property Fields Within High Resolution Models of the Gulf of Maine." Gulf of Maine 2050 Symposium. Portland, ME. November 4-8, 2019.

[20] Irving, J., McDowell, E., Pinckney, A., Ladue, T., He, R., Burkholder, K. (2019). "Modeling Subsurface Lagrangian Pathways in a Changing Gulf of Maine." Gulf of Maine 2050 Symposium. Portland, ME. November 4-8, 2019.