Improving regional cyberinfrastructure services through collaboration: Cyberteam for the Rocky Mountain Advanced Computing Consortium

Andrew J. Monaghan
Research Computing
University of Colorado Boulder
Boulder, CO, USA
andrew.monaghan@colorado.edu

Brett Milash
Center for High Perf. Computing
University of Utah
Salt Lake City, UT, USA
brett.milash@utah.edu

Tobin Magle
Steenbock Library
University of Wisconsin
Madison, WI, USA
tobin.magle@wisc.edu

Thomas Hauser
Research Computing
University of Colorado Boulder
Boulder, CO, USA
thomas.hauser@colorado.edu

Thomas E. Cheatham
Center for High Perf. Computing
University of Utah
Salt Lake City, UT, USA
tec3@utah.edu

Patrick J. Burns
Central IT and CSU Libraries
Colorado State University
Fort Collins, CO, USA
patrick.burns@colostate.edu

Shelley Knuth
Research Computing
University of Colorado Boulder
Boulder, CO, USA
shelley.knuth@colorado.edu

Anita M. Orendt
Center for High Perf. Computing
University of Utah
Salt Lake City, UT, USA
anita.orendt@utah.edu

Suzi White
Central IT and CSU Libraries
Colorado State University
Fort Collins, CO, USA
suzi.white@colostate.edu

Dawn Paschal
Central IT and CSU Libraries
Colorado State University
Fort Collins, CO, USA
dawn.paschal@colostate.edu

H.J. Siegel
Dept. of Electrical and Comp. Eng.
Colorado State University
Fort Collins, CO, USA
hj@colostate.edu

Timothy Kaiser
Acad. Net. and Comp. Resources
Colorado State University
Fort Collins, CO, USA
timk@colostate.edu

Andrew M. Johnson
University Libraries
University of Colorado Boulder
Boulder, CO, USA
andrew.m.johnson@colorado.edu

Becky Yeager
Research Computing
University of Colorado Boulder
Boulder, CO, USA
becky.yeager@colorado.edu

ABSTRACT

Advances in the volume, diversity, and complexity of research data and associated workflows requires enhanced capabilities to access, secure, reuse, process, analyze, understand, curate, share, and preserve data. To address this need at the regional level the University of Colorado, Colorado State University, and the University of Utah formed a "Cyberteam" in 2017 to provide cyberinfrastructure (CI) support to researchers at institutions in the Rocky Mountain Advanced Computing Consortium (RMACC) encompassing states across the Intermountain West. The Cyberteam is comprised of CI professionals across the three institutions who collaborate closely, sharing expertise and resources. Since its establishment, the Cyberteam has worked to broaden accessibility and options for computing, storage, and data publishing for RMACC researchers; enhance training on data- and workflow-oriented topics; improve engagement with researchers using CI; and better understand user needs and challenges. One key accomplishment has been the development of a series of focus group and survey instruments to achieve better understanding the CI needs and challenges of researchers across a diverse spectrum of disciplines. This paper provides an overview of the RMACC Cyberteam's objectives, accomplishments, challenges, and future direction.

CCS CONCEPTS

• Social and professional topics~Computing profession

KEYWORDS

cyberinfrastructure facilitation, high performance computing, workflows

1 Introduction

The Rocky Mountain Advanced Computing Consortium ("RMACC" or the "Consortium") has affiliated academic and government institutions across the Intermountain West. The Consortium was originally formed to focus on computation, data management, networking, and IT security, but has evolved into a robust regional collaboration supporting the ever-expanding areas of cyberinfrastructure (CI) and associated activities.

An NSF Campus Cyberinfrastructure award granted to three RMACC members -- the University of Colorado (CU), Colorado State University (CSU), and the University of Utah -- seeded the establishment of the RMACC "Cyberteam" in 2017 to enhance CI services for the entire Consortium. The establishment of the Cyberteam recognizes the critical need to provide "cradle to grave" data support as datasets continue to increase in size, diversity and complexity [1,2]. The goal of the Cyberteam is to improve capabilities to access, secure, reuse, process, analyze, understand, curate, share and preserve data for all RMACC institutions, including smaller entities that may have limited CI services onsite. This paper provides an overview of the RMACC Cyberteam's objectives, accomplishments, challenges, and future direction.

2 Methods

The Cyberteam is comprised of CI professionals across the three institutional grantees (CU, CSU, and Utah) who collaborate closely, sharing expertise and resources. The grant included seed funds to hire one CI professional at each university, each having a slightly different focus, and these members form the core of the Cyberteam. The CU professional supports the development of scientific workflows; the CSU professional supports data management and preservation; and the Utah professional supports data security, privacy and access controls for workflows. In practice, the work of each CI professional spans all of these areas.

The full Cyberteam communicates with each other and associated faculty staff through bi-weekly meetings conducted via Zoom video conferencing, and in-person at conferences and other meetings as opportunities arise. Individual team members communicate more frequently to address specific objectives of the Cyberteam and also do collaborative consulting with users for specific cases where inter-institutional expertise is required. The team maintains a Google Team Drive containing meeting minutes, presentations, lists of team tasks, reports, and other group-related documents.

3 Results

3.1 Broadening accessibility and options for computing, storage, and data publishing for RMACC researchers

An advanced high-performance computing (HPC) system, "RMACC Summit," [3] is shared by CU and CSU and housed at CU. RMACC Summit is a heterogeneous supercomputing cluster with ~12,500 cores primarily based on the Intel Xeon "Haswell" CPU. The nodes are provisioned with an Omni-Path Architecture interconnect which provides access to a 1.2 PB GPFS Parallel scratch filesystem. A 10% share of RMACC Summit is reserved for use by RMACC member institutions other than CU and CSU, a resource which is actively promoted by Cyberteam members. There are presently about 40 (non-CU/non-CSU) users from RMACC institutions including New Mexico State University, Boise State University, University of Utah, CU Denver, United States Geological Survey, and Idaho National Laboratory. The Cyberteam provides a limited volume of data storage at no cost to RMACC affiliates via CU's PetaLibrary (non-sensitive data) and Utah's Protected Environment (sensitive data). CSU's Mountain Scholar regional digital repository [4] provides a searchable data platform for data management and preservation of published scholarly and other works and includes an RMACC-specific collection into which RMACC affiliates can contribute at no cost. The RMACC Mountain Scholar collection presently has 240 entries.

Another means of broadening accessibility for RMACC researchers has been by containerizing software and workflows. Containers are a means of packaging software, scripts, and configuration parameters into a stand-alone object with a preferred environment that can be deployed and shared across platforms. They increase reproducibility and negate the need to recompile workflows when porting to different computing environments such as HPC clusters. Containers can be a particularly useful solution for migrating some research codes that may not compile easily outside of the Linux distribution within which they were developed, or without very specific compiler or library versions. Over the past year, Cyberteam CI facilitators have assisted users in developing containers for the following reasons: 1) because needed software packages were not available in the software stack on RMACC Summit or Utah's HPC clusters; 2) to employ legacy versions of basic software no longer supported in a given operating system; 3) to enhance the robustness of complex bioinformatics workflows that contain numerous packages and dependencies that often 'break' on HPC platforms due to shared library or software conflicts. Singularity is the containerization software deployed on RMACC Summit and on Utah's HPC clusters. Singularity is open-source software and well-suited for porting and scaling workflows for HPC. Additionally, Singularity can bootstrap containers from Docker, another widely-used containerization software.

3.2 Enhancing training on data- and workflow-oriented topics

The Cyberteam hosted several sessions at the 2018 RMACC HPC Symposium [5] surrounding the themes of data and workflow management. For example, three related "birds-of-a-feather" discussions were held on the topics of data life cycle planning, enhancing the reproducibility of data, and data preservation. Three hands-on tutorials were also hosted by the Cyberteam, including topics on integrating Globus into applications and workflows,

workflow management software, and using containers for scientific workflows in HPC environments. Collectively the sessions attracted about 75 participants, a strong indicator of the relevance of these topics. The Cyberteam will again host sessions at the 2019 RMACC HPC Symposium on several data and software topics [6]. Cyberteam members have taught a number of other data-, workflow-, and software-related courses at their respective institutions in addition to the RMACC symposium, and are developing new curriculum on topics including high throughput computing (HTC), Jupyter notebooks, and choosing the right computing environment for a given scientific workflow. A list of available trainings across the three Cyberteam institutions as well as other RMACC institutions is available at http://rmacc.org/trainings.

3.3 Improving engagement with researchers using CI

The hiring of CI facilitators in 2017 has enabled the Cyberteam to enhance CI user engagement by increasing consultation services. For example, in 2018 the CU CI facilitator participated in more than 60 user consultations; the Utah and CSU facilitators similarly conducted numerous consultations throughout the year. Topics included data publishing, data transfer, containerizing workflows, installation and optimization of software, parallelization, and HPC job submission. When appropriate, the facilitators referred consultees to resources such as the Open Science Grid, an HTC platform that supports large serial workflows.

Direct user engagement by Cyberteam CI facilitators has yielded substantial benefits in terms of improving the efficiency with which HPC resources are used. For example, the University of Utah CI facilitator worked with a user having performance issues with python code that processes weather forecast model data. The facilitator proposed several coding solutions to address I/O, memory, and processing bottlenecks that when implemented achieved a 40-fold increase in computational throughput. The CU CI facilitator consults with users when they are benchmarking workflows in preparation for large resource requests. This enables 'teaching moments' with researchers regarding best practices in compiling, linking, and scaling code; it is not uncommon for the resulting optimized workflows to be 50%-100% more efficient. While helping users optimize workflows is a standard aspect of HPC user support, the implementation of the Cyberteam has enabled enhanced support in this area.

3.4 Toward better understanding user needs and challenges

The Cyberteam is also engaging researchers by developing and conducting a series of focus groups and surveys aimed at better understanding and addressing the CI needs and challenges of researchers across a spectrum of disciplines. The blank questionnaires are published at osf.io/fdk43 [7] for other programs interested in using these as templates for their own user engagement efforts. The protocol for the focus group and survey instruments was submitted to the institutional review boards at each Cyberteam university and approved with 'exempt' status.

The first instrument is a ~90-minute in-person semi-structured focus group discussion intended to collect open-ended information on research groups' cyberinfrastructure usage including data management and preservation, collaborators, remote science activities, use of cloud computing services, data transfer, workflows, software use, use of sensitive data, data sharing, access to support, future needs, and outstanding issues. The focus groups are being conducted throughout 2018-2019 with the goal that each of the three Cyberteam institutions will conduct 10-15 interviews with research groups spanning a broad range of disciplines from the physical, social, and biological sciences and humanities. Results thus far indicate common challenges for research groups are the need for software modernization to take advantage of today's HPC environments, and a need for training on best practices for data storage and preservation. Numerous issues were solved via simple communication of systems, services, workflows, and structures already in place at the institutions, of which the researchers were unaware. One of the most significant lessons was this "awareness gap," that only became apparent when Cyberteam staff visited the researchers - they are typically extremely independent, and are not prone to requesting advice and support from the Cyberteam. Cyberteam members have and will follow up with each group summarizing some of the key challenges they face and offering to help address these to the extent possible.

A short follow-on survey will be given to the focus group participants electronically via Qualtrics about one year after the interviews to assess whether the focus group discussion and subsequent engagement by the Cyberteam changed CI usage among group members. Finally, a longer format questionnaire on CI usage covering the same topics discussed in the focus group interviews will be broadly distributed to CI users across the universities to obtain a larger and more diverse sample of responses. We anticipate that the results of the focus groups and surveys will help us improve CI services and training and plan for the future CI needs of RMACC researchers.

4 Discussion and Conclusions

The RMACC Cyberteam was formed by CU, CSU, and the University of Utah in 2017 to address the CI needs of researchers at institutions in the Intermountain West. The Cyberteam is building a regional community of CI professionals that supports their professional development and facilitates sharing of expertise and resources. In a short period of time the Cyberteam has been successful at broadening accessibility to campus-level and community-level CI resources, building institutional and regional CI knowledge capacity, enhancing training opportunities, expanding user engagement, and improving workflow efficiency. The principal investigators at the three grantee universities are communicating these successes and articulating the value of the CI facilitators to their leadership to secure sustained institutional

funding for the positions beyond the end of the NSF grant. All three institutions are enthusiastic about continuing the Cyberteam activities beyond the end of the grant. These activities include having regular team meetings to discuss common issues, hosting joint sessions and trainings at regional conferences, and when appropriate working on collaborative proposals that benefit the RMACC community.

Some key challenges identified by CI facilitators in the first years of the Cyberteam have included meeting the high demand for personalized consultation services to support users' unique data and workflow needs, and changing behaviors with respect to data management and publishing best practices. Objectives of the Cyberteam in the forthcoming year are to finish conducting the focus groups and surveys and synthesize the results into actionable information, and to increase outreach and engagement with smaller RMACC institutions to raise awareness of resources and training opportunities.

ACKNOWLEDGMENTS

This work is supported by the NSF Campus Cyberinfrastructure Program (Award # 1659425). RMACC Summit is supported in part by the National Science Foundation under awards ACI-1532235 and ACI-1532236. The survey instruments were approved under Protocol 18-0101 (CU) and Protocol 234-18H (CSU) for review category "Exempt – Category 2." IRB approval is pending for Utah (Protocol IRB_00113715).

REFERENCES

[1] Heidorn, P.B., 2008. Shedding light on the dark data in the long tail of science. Library Trends, 57(2), pp.280–299.

[2] Ogburn, J.L., 2010. The imperative for data curation. portal: Libraries and the Academy, 10(2), pp.241–246.

[3] Anderson, J., P.J. Burns, D. Milroy, P. Ruprecht, T. Hauser and H.J Siegel, 2017. Deploying RMACC Summit: an HPC resource for the Rocky Mountain region. In Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact (p. 8). ACM.

[4] Colorado State University (CSU), 2019. Mountain Scholar Digital Collections of Colorado and Wyoming [cited 10 April 2019]. Available from: https://mountainscholar.org.

[5] Rocky Mountain Advanced Computing Consortium (RMACC), 2018. RMACC 2018 HPC Symposium schedule [cited 10 April 2019]. Available from: https://rmacc2018hpcsymposium.sched.com.

[6] Rocky Mountain Advanced Computing Consortium (RMACC), 2019. RMACC 2019 HPC Symposium schedule [cited 10 April 2019]. Available from: https://rmacc201ihpcsymposium.sched.com.

[7] Magle, T., B. Milash, A. Monaghan, T. Hauser, T. Cheatham, P. Burns, A. Orendt, S. Knuth and A. Johnson, 2019. RMACC Cyberteam surveys. Open Science Framework. osf.io/fdk43. DOI 10.17605/OSF.IO/FDK43.

.