Connecting the Non-Traditional User-Community to the National CyberInfrastructure

Ritu Arora  
Texas Advanced Computing Center  
The University of Texas at Austin  
Austin, Texas 78758  
rauta@tacc.utexas.edu

Jessica Trelogan  
Institute of Classical Archaeology  
The University of Texas at Austin  
Austin, Texas 78759  
j.trelogan@austin.utexas.edu

ABSTRACT
This paper reports on a hands-on workshop that was organized to promote the usage of the National CyberInfrastructure (NCI) amongst non-traditional High Performance Computing (HPC) users. The majority of the workshop participants were students and professionals who had never used the NCI before but were interested in leveraging it for conducting computationally intensive Big Data management activities. With the support from the National Science Foundation (NSF), students from underrepresented groups were also funded to participate in the workshop, where they learnt about both the NCI and Big Data management. The workshop itself was an outcome of an XSEDE Extended Collaborative Support Service (ECSS) project that involved non-traditional HPC users from the archaeology domain.

Categories and Subject Descriptors
D.1.3 [Programming Techniques]: Concurrent Programming - distributed programming, parallel programming; K.3.2 [Computer and Information Science Education]: Computer science education.

General Terms
Measurement, Performance

Keywords
High Performance Computing; parallel programming; Big Data; national cyberinfrastructure

1. INTRODUCTION
The need for training current and future generations of professionals in large dataset management within the NCI is increasing in direct proportion with advancement in disciplines ranging from science and technology to the humanities and social sciences. Curators at the Institute of Classical Archaeology (ICA) at the University of Texas at Austin are one example of non-traditional users leveraging the NCI for managing a large data collection. The ICA data collection is actively evolving, but has a long history, spanning several decades, and contains a wide variety of data types. Approximately 4 TBs in size at the time of the workshop, it was in a disorganized state, had a deeply nested directory structure, and had become unwieldy for ongoing research and publication tasks. Common data management tasks such as searching, identifying, sorting, deduplication, and reorganizing were, for the most part, conducted manually by the curators on personal computers connected to a local file share. These tasks were burdensome, time-consuming, and overwhelming due to the complex and evolving nature of the collection.

There was not only a pressing need for a large and secure storage space to house the collection and to ensure its preservation, but also for automating routine data curation tasks including the creation of descriptive metadata, reorganizing files, updating file formats, and identifying redundant data. To analyze the contents of the collection, file format identifications were extracted using DROID [1], and file checksums were calculated via a set of custom scripts to aid in deduplication. However, the curators had restricted access to the server on which the data was originally kept and could thus only run DROID themselves over a slow network connection. Even with help from a system administrator with direct access to the server, it took more than two days to extract the metadata from a subset of the data collection. Because the data collection was in active use while legacy data were being integrated with it, frequent snapshots of the data collection were required to iterate through data curation tasks. Hence, large-scale storage and processing resources were explored for conducting the data management tasks frequently, efficiently and in a timely manner.

The storage space for the project was provided on the resources at the Texas Advanced Computing Center (TACC) [2] and an XSEDE [3] allocation along with ECSS [4] support was requested for the project. Because ICA curators were inexperienced with large-scale HPC resources, training was required as a first-step of the ECSS project. Basic instruction was provided in the use of remote HPC platforms like Stampede, along with introductory Linux training, and an appraisal of various system-related policies. Later in the project, a semi-automated metadata extraction workflow was set up on Stampede with the help of ECSS staff to enable ICA curators to perform computationally intensive data curation activities. By parallelizing the steps in their workflow, the time required for metadata extraction and checksum calculation was drastically reduced from days to just a few hours. The curators were then trained to perform these workflow steps independently on Stampede. By the end of the ECSS project, the data curators, with no prior exposure to large-scale HPC platforms, were enabled to use Stampede for their routine data management tasks [5], dramatically improving their ability to manage their own collection without interrupting their main research and publication activities.

The ECSS project was a success. It not only helped ICA meet its preservation and data curation needs, it also brought to light a number of challenges that may act as a barrier for non-traditional HPC users to make use of the NCI. With guidance and training from the outset, users with no prior experience can make great strides toward overcoming knowledge barriers and leverage the power of HPC. Details that may seem obvious to a more experienced user, including choosing appropriate data transfer protocols and navigating user environments of remote HPC resources cannot be overlooked. For the ICA team, a detailed
training document was developed so they could perform curation workflows on a remote HPC platform like Stampede and use it to train others in their team. This stakeholder training aspect of the XSEDE project, including the development of the training documentation, led to the workshop described in this paper. The idea was to conduct the workshop at an international conference in order to connect with a wider community of student and non-traditional HPC users and share our experiences with other potential users. The rest of this paper presents a report on this workshop.

2. OVERVIEW OF THE WORKSHOP
Sustainable, well-organized and well-documented data collections are critical for making discoveries through data-driven and data-intensive research. For preparing such a data collection, in addition to efficient and scalable data management tools and techniques, well-trained data curators and data managers are essential. Data management includes a variety of tasks, *viz.* data transfer, checking data integrity, and data preservation. At many organizations, despite the rapid growth in the size and complexity of the datasets, such data management tasks are still being conducted on desktop computers and single-node servers. The hardware and software limitations of these resources make it difficult to conduct routine data management activities efficiently for large datasets. Therefore, it is imperative to leverage HPC and massive storage resources for timely processing and management of large datasets. Even though such resources are available to data curators and data managers through a NCI like XSEDE without entailment any direct cost, the learning curve associated with leveraging remote supercomputing resources poses a significant adoption barrier. The learning curve starts at the level of understanding of the Linux operating system, and the user-environment of the remote supercomputers. These barriers and challenges were noticed first-hand as part of the ECSS project described above, in Section 1.

The “First Hands-On Workshop on Leveraging High Performance Computing Resources for Managing Large Datasets” [6] was conducted specifically to lower the adoption barriers to scalable high-end resources that are available to the research community through the NCI. The workshop was held on October 27, 2014 at Washington DC and was collocated with the 2014 IEEE International Conference on Big Data [7]. The workshop was attended by students and professionals from a wide range of disciplines, including: library and information science, computer science, mental health, electrical and computer engineering, and biology. There were more than 45 participants at the workshop during the sessions.

At the workshop, the participants were (1) introduced to scalable High Performance Computing (HPC) and high-end storage resources that are available through the NCI, (2) introduced to the user-environment and policies of these remote HPC platforms, (3) trained in transferring data to and from the remote HPC platforms, (4) trained at installing software on remote HPC platforms, and (5) provided hands-on experience in extracting metadata from a sample data collection on a remote HPC platform. The participants were also introduced to off-the-shelf tools for information visualization and their significance in assessing the data collection. Participants were not required to have any prior knowledge about HPC and were provided accounts for accessing TACC and NERSC [8] resources from their laptops. A detailed agenda of the workshop and the presentation material are available at [9].

The workshop succeeded in engaging non-traditional HPC users (*e.g.*, data curators and librarians) in leveraging the NCI for Big Data management activities. *Such non-traditional HPC users need to be trained in leveraging the NCI efficiently for addressing the management of large data sets in various scientific and engineering disciplines.*

In addition to the hands-on sessions on HPC resources for Big Data management, the workshop agenda included invited talks from practitioners in the areas of HPC and Big Data. The speakers discussed the challenges and opportunities that are at the cusp of HPC and Big Data. They presented test cases from genomics, archaeology, astronomy, fraud-detection, and healthcare domains. An improvscience [10] session was organized to serve as an ice-breaker between participants, who then formed small teams to solve problems related to data management. These real-world problems were provided to them as challenges during a “hackathon” session towards the end of the workshop. The scope of the problems covered the length and breadth of the material presented at the workshop.

The workshop was enriched by the inclusion of curious students from diverse backgrounds. With support from the National Science Foundation (NSF), 20 students from underrepresented groups were offered travel grants to attend the workshop and to participate in the 2014 IEEE International Conference on Big Data. Out of these, one declined to participate just few days prior to the workshop due to the heavy load of course-work. During the student selection process, consideration was given to applicants from groups that have been traditionally underrepresented in HPC and Big Data, *viz.* women, African-Americans, Hispanics, Native Americans, Alaska Natives, Pacific Islanders and people with disabilities. Table 1 shows the underrepresented groups to which the selected applicants belonged. Out of the 20 selected students, 18 were female and two were male, both of Hispanic descent. Out of the 18 female students selected, two were of African-American descent, two were of Hispanic descent, and one was physically disabled. The travel grant covered air-fare, hotel, workshop/conference registration fees, ground transportation, and per diem expenses for the selected students. In addition to NSF, other sponsors of the workshop were TACC, NERSC, Lawrence Livermore National Lab (LLNL), and XSEDE.

<table>
<thead>
<tr>
<th>Underrepresented Group</th>
<th>Number of Selected Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women</td>
<td>13</td>
</tr>
<tr>
<td>Women, and African-American</td>
<td>2</td>
</tr>
<tr>
<td>Women, and Hispanic/Latino</td>
<td>2</td>
</tr>
<tr>
<td>Women, and Physically Disabled</td>
<td>1</td>
</tr>
<tr>
<td>Hispanic/Latino, Male</td>
<td>2</td>
</tr>
</tbody>
</table>

3. PRESENTATIONS & INVITED TALKS
There were four invited speakers at the workshop who presented Big Data challenges from various domains and helped in setting the context for the hands-on sessions.

The first speaker was invited from the Department of Energy Joint Genome Institute (JGI). She presented an overview of the HPC resources at the JGI and then talked about the TeraBytes of genomic sequence data that are generated per day at their facility.
The speaker then provided an overview of the challenges in managing such large volumes of genomic data during and after processing at the JGI. This was followed by an overview of their system named as JGI Archive and Metadata Organizer (JAMO), and their data management policies.

The second invited speaker was from ICA at the University of Texas at Austin. She spoke about the challenges—discussed above in Section I—that she and other curators at ICA have faced in managing archaeological data collected over several years. She then talked about how HPC and the ECSS support through XSEDE enabled her team to iteratively assess their complex data collection as it continues to evolve, in a timely and efficient manner. She highlighted the importance of training and documentation to overcome some of the technical challenges that non-traditional HPC users face while using the NCI.

The third invited speaker was from the International Data Corporation, a market research company that has been analyzing the trends in HPC and Big Data for many years. He presented an overview of how Big Data is redefining computing and the HPC space. He presented a summary of some of the major challenges and market trends in using Big Data today. Then, he presented some real-world examples of Big Data from various scientific and business domains, and spoke about how HPC and high throughput computing are enabling data-intensive computing across these domains. Some of the examples that he presented included: how HPC is being used for fraud detection by the United States Postal Service, for real-time insurance quotes by GEICO, and for fraud detection by PayPal.

The fourth invited speaker was from the Globus [11] team at the University of Chicago. He gave a demo of the Globus Connect Personal tool that can be used for sharing and transferring large volumes of data through a user-friendly interface.

An overview of XSEDE and TACC resources was provided to the workshop participants. The process of requesting allocation on XSEDE resources was discussed and the participants were apprised about the policies and practices at the open-science data centers.

4. HANDS-ON SESSIONS
The workshop participants were provided temporary accounts for using the TACC and NERSC resources. They connected to the Stampede supercomputer at TACC and/or to the Hopper supercomputer at NERSC. Then they learnt about the basic Linux commands and familiarized themselves with the user environment of the remote supercomputers. As the majority of the workshop participants were trying Linux for the first-time and as there were a limited number of mentors/instructors that help them at the workshop, it was challenging to bring them up to speed in a very short span of time. Some of the workshop participants, who had backgrounds in computer science or engineering, were recruited ahead of the workshop to assist their fellow participants during the hands-on sessions at the workshop and this proved to be very helpful.

The workshop participants were provided a heterogeneous sample dataset to use during the exercises on metadata extraction and checksum calculation. They learnt about the basic protocols for file-transfer and practiced transferring data to/from Stampede and/or Hopper supercomputers.

They were then given instructions on downloading and installing DROID, a tool for metadata extraction and file format identification. While most of the participants were able to download DROID, several ran into challenges with DROID installation. Not all the participants were able to complete the exercises using DROID for metadata extraction of the given dataset, but all were exposed to the process, understood the basic concept, and were provided with clear instructions on how to proceed at their leisure.

There were some lessons learnt from the challenges faced during the hands-on sessions. For one, either such an intensive workshop should be spread across two days to allow sufficient time for participants of all skill levels to keep pace with instruction, or the content covered during the workshop should be reduced. One option would be to offer the workshop in a tutorial mode in future so that sufficient time can be devoted to the hands-on sessions. Other challenges related to the configuration of the conference room in which the workshop was conducted. In future, a workshop-style room, with tables, electrical outlets and good network connectivity would be desirable. This session was conducted in a room configured for lectures, making hands-on work more difficult.

5. IMPROVISATION AND HACKATHON
An improvisation session was organized to serve as an ice-breaker amongst the workshop participants and to help them get organized into small teams to work on the challenges given to them during the hackathon session.

The participants were randomly organized into small teams, each of which participated in improvisational games and exercises. This session promoted communication amongst the team members and made them comfortable working with each other during the hackathon session.

During the hackathon session, the teams were provided four challenging problems that were modelled on the basis of some real-world scenarios found in the HPC and Big Data domains. The hints and knowledge required to solve the challenges were imparted during the workshop sessions prior to the hackathon. Each team had to select one challenge out of the four given to them, and develop a solution strategy for it. Each team was then asked to prepare a presentation on the challenge they selected and the solutions they proposed.

The hackathon was motivated by the fact that, in the real-world, large multi-disciplinary computational projects are often completed with the collaboration of multiple stakeholders. The wide range of perspectives and experiences of different stakeholders impacts the design of solution strategies. Developing trust and cooperation amongst team members is critical in accomplishing project objectives in a timely fashion and in meeting stakeholder needs. The workshop participants had diverse skill-sets and backgrounds. When they formed small teams to work on specific problems, they not only got a chance to understand different perspectives to problem-solving, but also gained experience in optimally utilizing the skills of their team-members.

6. MENTORING ACTIVITIES
Mentoring was provided to the funded students at the workshop both during and after the workshop. The funded students were organized into groups during the lunch so that they could have one-on-one conversations with invited speakers, and additional professionals at the workshop. They discussed technical and
career-related questions. A networking dinner was also organized for the students to interact with the recruiters from national laboratories and discuss career options with them. A LinkedIn.com group was set up to encourage communication amongst the students and other mentors in the group. Information about career and other opportunities is posted on this group to keep the students informed about additional opportunities in the areas of HPC and Big Data.

7. SURVEY AND FEEDBACK
Eighteen workshop participants responded to the post-workshop survey conducted through SurveyMonkey.com, and snippets of some of the responses is presented in this section. As can be seen from the data presented in Figures 1-5, the majority of the respondents to the survey felt that the workshop was useful to them in one way or the other.

As evidenced from the snippet of the responses presented in Figure 5, it was encouraging to see that all the participants reported that they learnt at least something after attending the workshop. About 94.44% of the respondents felt that the workshop provided enough information to further explore the TACC and XSEDE resources. The data presented in this paper, the responses to the additional questions on the survey, and the additional feedback provided through the email correspondence indicate that the workshop succeeded in achieving most of its goals. The feedback provided by the participants will be used to improve the quality of future workshops. For example, one participant was disappointed at not seeing the topic of databases covered at the workshop. Databases are an important topic and effort will be made to cover it during future workshops.

![Figure 1. Rating of the Hackathon session](image1)

**Figure 1. Rating of the Hackathon session**

![Figure 2. Rating of Workshop Meeting the Expectations](image2)

**Figure 2. Rating of Workshop Meeting the Expectations**

![Figure 3. Overall Experience at the Workshop](image3)

**Figure 3. Overall Experience at the Workshop**

![Figure 4. 94.44% of the respondents felt that the workshop provided enough information to explore TACC/XSEDE resources](image4)

**Figure 4. 94.44% of the respondents felt that the workshop provided enough information to explore TACC/XSEDE resources**

8. CONCLUSION
The workshop reported in this paper succeeded in enabling non-traditional HPC users, including students from underrepresented groups, in accessing supercomputers like Stampede and Hopper, and in developing an understanding about the NCI. The workshop participants developed an insight into some of the challenges and possibilities that the Big Data era brings and how they can be solved and leveraged using HPC. As per the post-workshop survey and feedback, the workshop participants would be interested in attending the second workshop on this topic.

Additional workshops along the lines of the workshop reported in this paper can be instrumental in engaging non-traditional user-communities to make use of the NCI. However, the format of the workshop needs to be modified in future to address some of the concerns raised by the participants. An example of a concern raised...
was that there was not enough time to finish the exercises given during the hands-on session. Such concerns can be addressed by spreading the workshop over two-days or by reducing the amount of activities and topic covered during the workshop.

**Figure 5. Learning at the Workshop**

Given the cost- and time-constraints in organizing such workshops, organizing one such workshop every year seems feasible. The needs of non-traditional HPC users can be better understood through carefully designed workshop sessions (viz. hackathon). Such needs can be factored in during the design of future HPC platforms that become the part of the NCI and the associated usage policies.

While it was expensive to collocate the workshop with an international conference like 2014 IEEE International Conference on Big Data, an advantage of doing so was to connect with the user communities that we would not have reached otherwise. The students who were supported by the NSF grant to participate in the workshop also got a chance to participate in various other sessions at the conference and to network with attendees outside the workshop.

In general, if a workshop is a part of the conference, and hence does not have dedicated registration process for it, it is hard to get the exact number of the participants *a priori*. If the information about the participants can be obtained before-hand then the process of distributing the training accounts on the supercomputers can be streamlined. Besides this, having an idea about the approximate number of participants in advance will be useful in planning and requesting support from other colleagues during the hands-on sessions.

On the basis of the workshop experience, the authors highly recommend such hands-on and interactive workshops as a means of engaging non-traditional HPC users. The authors also note that the availability of funds for sponsoring the participants from the underrepresented groups to attend such workshops is critical for their extreme and sustained inclusion in the usage of the NCI.

9. **ACKNOWLEDGEMENT**

We are grateful to NSF for sponsoring the workshop through the NSF grant number 1447000. We are also grateful to colleagues from UT Austin (Allison Warner, Hedda Frochaska, Faith Singer-Villalobos, Katie Cohen, Maria Esteva, Melyssa Fratkin, Niall Gaffney, Paula Baker, Rosalia Gomez, and Valerie Shilling) for their various contributions in organizing the workshop. We are grateful to the user services group at TACC and staff at NERSC for preparing the training accounts on Stampede and Hopper. We thank the workshop committee, invited speakers, sponsors and participants for their contributions in making the workshop a success.

10. **REFERENCES**


