Training-based Workforce Development in Advanced Computing for Research and Education (ACoRE)

Semir Sarajlic*[1], Neranjan Edirisinghe*[2], Yubao Wu[3], Yi Jiang[4], Gregori Faroux[5]

ABSTRACT

Data proliferation across all domains is enabled by the advancement in data sharing and procurement that has led to explosion in volume of data worldwide. The exponential growth in data has provided opportunities for machine learning algorithms and computational algorithms that benefit from greater availability of data. Access to robust and resilient high performance computing (HPC) resources is available through campus and/or national centers. The growth in demand for advanced computing resources from non-traditional HPC research community leads to an increasing strain on user-support at local centers and/or national programs such as XSEDE, which creates challenges. At Georgia State University (GSU), we have a diverse HPC user community of over 300 users whom represent more than 40 disciplines and use about four million CPU hours annually. We implemented a workforce development model to support our HPC community through a combination of local and external workshops as well as integration of HPC training within classes such as Big Data Programming, Scientific Computing, Parallel and Distributed Computing (PDC). As part of our approach, we consolidated our three disparate local HPC resources under one management system that provides a similar user environment to that of national resources part of XSEDE, which simplifies user transition from local to national resources.

CCS Concepts

• **General and reference~Metrics** • *General and reference~Design* • **Applied computing~Education** • Computer systems organization~Heterogeneous (hybrid) systems

Keywords

Workforce development, training, High Performance Computing, sustainability, metrics

1. INTRODUCTION

Data proliferation across all domains is enabled by the advancements in data sharing and procurement that has led to an explosion in the volume of data worldwide that rapidly grows at an exponential growth [1]. This explosion of data is driven by automated instruments, social media, and scientific research [2-4].

*Corresponding Authors.

[1] Research Solutions, Georgia State University, ssarajlic1@gsu.edu
[2] Research Solutions, Georgia State University, neranjan@gsu.edu
[3] Computer Science, Georgia State University, jwu28@gsu.edu
[4] Mathematics and Statistics, Georgia State University, yjiang@gsu.edu
[5] Research Solutions, Georgia State University, gfaroux@gsu.edu

PEARC17, July 09-13, 2017, New Orleans, LA, USA

ACM ISBN 978-1-4503-5272-7/17/07.
http://dx.doi.org/10.1145/3093338.3104178

In dealing with the demanding computing capabilities to process and analyze the data, parallel and distributed computing (PDC) now permeates most computing activities. The rapid growth in importance in data science, machine learning, and deep learning, are making most users dependent on advanced cyber-infrastructure for parallel processing and advanced data analytic techniques. As a result, we face a multifaceted challenge in dealing with the deluge of data, and significant challenges in transforming the curriculums across domains to prepare the students to enter the workforce. Recognizing the challenges, in 2015, the White House presented an executive order to create a national strategic computing initiative that addresses the cyber-infrastructure capabilities and workforce development that enable the nation to maintain its competitive advantage [5]. Advanced Cyber-Infrastructure Research and Education Facilitator (ACI-REF) program [6] and XSEDE's Campus Champion program [7] are national programs that help with the facilitation of Cyber-infrastructure resources in research activities across campuses. Virtual Residency program [8] provides education and training for the Facilitators. At university curriculum scale, a gap in the educational curriculum for emerging technology focusing on the parallel and distributed computing paradigm was identified, and to address this rapid change in technology, Georgia State University lead a collaborative project "Parallel and Distributed Computing Curriculum Development and Educational Resources" ($1, 100, 707 – 09/01/2012 – 08/31/2017) [9, 10]. The success of this project and its concept inspired our approach in our workforce development model for promoting and facilitating research and training activities for our HPC community. In the following section, we present our workforce development model for GSU, and we describe the strengths and weaknesses of some of our initiatives as part of the model.

2. WORKFORCE DEVELOPMENT MODEL

Our objective was to form a sustainable HPC workforce development model for supporting our growing HPC community that empowers our users and enables them to disseminate gained knowledge to their peers. Our model (depicted by Figure 1) is broken into two building blocks: education and cyber-infrastructure (CI). Education block is made up of community engagement activities such as workshops, open-door sessions, wiki/slack, annual symposium, community feedback, and coordination with faculty in integration of HPC into their course(s). CI block acts as a foundation for our model as our educational activities rely on the ability for our users too readily and seamlessly access HPC resources. CI block focuses on three practices, which are simplify access to HPC resources, simplify HPC usage, and consolidate HPC resources. The following sections will cover the details of the approach and the outcomes of the activities.

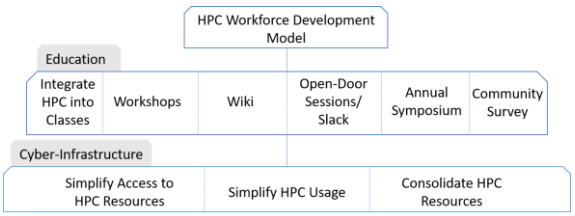

Figure 1. HPC workforce development model at Georgia State. Our educational model relies on strategically deployed HPC resources along with services that enable readily access to HPC resources.

3. APPROACH

3.1 Integrate HPC into Classes

3.1.1 CDER

CDER is an NSF sponsored cluster with a purpose of enabling the efforts in addressing the gap in the education for the emerging technology focusing on the Parallel and Distributed Computing (PDC) paradigm and the missing curriculum in the education to address this rapid change in technology. The PIs (Sushil K. Prasad, Anshul Gupta, Arnold Rosenberg, Alan Sussman, and Charles Weems) addressed these challenges via a project devoted to creating and sustaining curriculum and educational infrastructure to facilitate the teaching of PDC topics in undergraduate computer-related curricula. The goal of the project is for every graduating student to become skilled in PDC technology, hence be prepared to enter tomorrow's workforce [9, 10]. The success of this project inspired our effort to take a similar approach in integrating institutional supported HPC resources into classes that is described in the next section with our DICE HPC resource.

3.1.2 DICE

In providing research computing services at GSU, we discovered that most of the students who conduct computational research using HPC resources did not receive training in their classes to use these resources. This was more common with students representing *long tail of science* [11] as they do not have access to computational classes within their program. As results, PIs Edirisinghe and Sarajlic proposed a student dedicated cluster (Data Intensive Computing Environment – DICE) that would serve as a platform to learn the Big Data analytic methods and develop their research on the cluster. In Spring 2017, DICE was integrated into Big Data Programing and Scientific Computing classes. This approach enables students to test and develop their class work on a production cluster.

3.1.2.1 Big Data Programming Class

Big Data Programming is a newly created class at GSU. It covers the most commonly used frameworks and tools including MapReduce/Hadoop and Spark. Students are required to do hands-on projects to practice programming in Hadoop and Spark.

In the classes, the students first program in a local single-node cluster like Cloudera or Ubuntu where Hadoop and Spark are installed. Students can debug and verify their algorithms on small datasets. If it is successful, then they submit jobs on DICE, the real cluster to run the job on a large dataset. This local-to-real-cluster workflow model is widely adapted in many companies. The experience will benefit the students when they are in the job market.

The class benefited from DICE's friendly user interface with diverse functionalities, which is supported by Jupyterhub and Ambari. Students were able to run Hadoop or Spark in different programming languages including Java, Python, R, and Scala. In addition to the cluster, students had access to a documentation knowledgebase and a user forum via Slack where students could ask questions and discuss with other students (see section 3.3). Through discussion, the students can resolve their problems and learn more quickly.

3.1.2.2 Scientific Computing with MATLAB Class

Scientific Computing is a new class for undergraduate students in Mathematics and Statistics (Math & Stat) at GSU. The goal of the class is to help the students understand the basic concepts of programming and computing, gain working knowledge of scripting / programming. We choose MATLAB for computing, data processing and visualization. Students learn useful functions from linear algebra (e.g. matrix inversion, eigenvalues), differential equations (ODEs), symbolic calculations and statistics. Students were exposed to Image Processing, Signal Processing and Bioinformatics toolboxes. In addition to undergraduate students from Math & Stat, and Computer Science, several graduate students from Math & Stat, Biology, and Neuroscience Institute also audit the classes. The latter belong to the *long tail* that did not receive any programing training and found this class useful. In the classes, all lecturers are given as ipynb on DICE. The students have easy access to the code and execution. Their homework and projects are all submitted to the DICE server in iphynb format. Students who miss a few lectures can easily pick up the progress on their own using the ipynb lecture notes.

3.2 Workshops

To facilitate adoption and use of HPC resources, we conduct periodic workshops such as Introduction to Workload Management in HPC Environment, Introduction to Fundamentals in Data Science, and XSEDE's monthly HPC workshops. Workshops have proven to be a preferred medium for learning the HPC techniques among students and researchers representing *long tail of science* who do not have access to computational classes in their program. All the workshops feature both lecture and hands on components that maximizes attendees' reception of the material.

3.2.1 The Fundamentals in Data Science workshops

We initiated "The Fundamentals in Data Science workshop series" that focuses on introducing data science fundamentals using Python language to boost our workforce in data science. Python language was our first choice due to its versatility. Our audience is mainly representing *long tail of science* for whom Python provides a middle ground in terms of programming complexity, usability, and productivity. Furthermore, Python provides all-in-one solution (i.e., users can learn one language and use it for most of their needs). This workshop series consist of five workshops, four hours each, during which we cover, introduction to python and numpy, Scipy and matplotlib, Pandas, Machine learning with scikit learn and deep learning with Tensorflow. Each workshop was heavy on hands-on exercises using real life openly available data. Our goal was to provide workshop participants an experience where they interact with real-world situations while instructors are available for assistance whenever needed. We encourage participants for knowledge exploration using open source communities, such as google groups, stack overflow etc. Our involvement as instructors was to introduce the subject and guide them on how to find solutions for their problems using self-learning. Authors believe such training is necessary in navigating increasingly fluid and complex technology landscape.

3.3 Open-Door Sessions, Wiki and Slack

Open-door sessions were facilitated in Fall 2015 to provide researchers an opportunity to come and meet the facilitators informally and talk about their technology pains. The attendance for the open-door sessions was low throughout the Fall term with the attendance peaking during the launch of Orion in October 2015. Since then we have relied more on the Wiki (i.e., knowledge base) and Slack as a forum.

A knowledge base containing the documentation of HPC resources provides users easy access to information on getting started with the resources. Furthermore, knowledge base features integration with ticketing system that enables users to report bugs and request features (e.g., application/package installs). We utilized Slack for a forum for students to ask questions and discuss with other people. Through the discussion, students can resolve their problems and learn more quickly.

3.4 Annual Symposium

Scientific Computing Day (SCD) Symposium [12] provides a venue for students and junior researchers to share their work with the university community and exchange views on multidisciplinary computational challenges and current developments. SCD has become a focal point for the research computing community at Georgia State. Since the inauguration in 2015, the success of SCD has been exceptional with attendance more than doubling for SCD '16 (Figure 2).

SCD is a highlight of our workforce development efforts, for example, we host monthly workshops that are discipline specific and XSEDE affiliated workshops throughout the year that benefit the students, and

in part is the reason we had a stellar growth in our student poster presentations (2015 – 11 posters, 2016 – 26 posters). SCD is a measure of our growth and success as a research computing community at Georgia State, and SCD acts as a highlight of all our workforce development efforts that take place throughout the year. More importantly, SCD is the time that we all meet as a community and share our work with one another and exchange views on multidisciplinary computational challenges, current developments, and needs assessments.

3.5 Community Survey

In order to facilitate adoption and use of HPC resources, we surveyed our community in order to assess the current needs and technology challenges our researchers encounter. Feedback from the community is critical to planning new initiatives. In February 2017, we surveyed our research community of 1,447 current/past PIs with a custom internally created Qualtrics survey with the theme question: "What technology challenges do you have that are impacting your current research progress?" We allowed the PIs to select multiple answers for this question to give us a better understanding of the challenges PIs face. Then, we provided specific questions in relation to the responses in the theme question. About 8% of the researchers responded to the survey, in which one of the key challenges was the lack of technical training.

3.6 Simplify Access and Usage of HPC

Our approach in facilitating HPC workforce development model relies on implementation and deployment of readily accessible HPC resources, which defines our model's foundation. As previously reported by authors Sarajlic et.al. in [13], access to HPC resources is simplified by an automated portal for provisioning user accounts, which enables users to rapidly gain access to the resources. Furthermore, authors reported on their technique in simplifying HPC usage and empowering users to utilize the resources by developing wrappers (i.e., automated job submission scripts) for common open-source and proprietary applications such as R, Python, MATLAB®, STATA®, Gaussian 09, SAS®, and Trinity.

3.7 Consolidate HPC Resources

Disparate clusters utilizing different workload managers create a challenge for facilitators in supporting non-uniform systems with a rapidly growing research computing community with majority of them being non-traditional users representing the *long-tail science*. Additionally, disparate clusters create a challenge in supporting users and facilitating training activities as well as maintaining documentation/wiki. As a result, our approach was to consolidate clusters under one management platform with SLURM as our preferred workload manager (Figure 3). Our past/current clusters used Load Leveler, IBM's LSF, SGE, and Open Lava (Open-Source LSF). With our consolidation strategy, SLURM was a preferred choice as it is open-source, robust, and coincides with GSU's HPC model, which relies on national resources in meeting the demand of our select users with large HPC/HTC workloads [13, 14]. Figure 3 illustrates our modular/condominium approach in integrating clusters Orion, DICE, and CDER under one management platform, ACoRE – Advanced Computing for Research and Education.

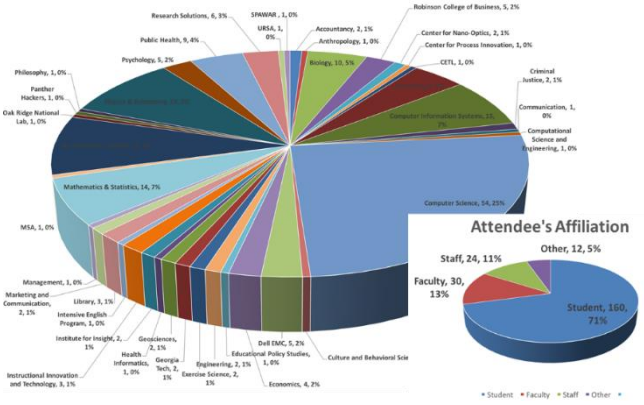

Figure 2. SCD16 received over 220 registrants (165 attendees), whom represented 35 departments/disciplines. Over 60 authors contributed to the presentations, which included 27 poster presentations. SCD16 featured partners and special guests from United States Navy's SPAWAR, DELL EMC, South Big Data Hub, and Georgia Tech. [12]

4. RESULTS AND DISCUSSION

In previous sections, we presented our workforce development model that addressed our community engagement activities along with our cyber-infrastructure deployment strategy. Our efforts in integrating HPC resources into curriculum to provide student with a production cluster to test and develop their research is enabling us at institutional level to train and prepare students for the Data Science field. Additionally, workshops are the preferred medium for non-traditional researchers and students who do not have access to computational classes within their program but need to learn the basic techniques to achieve their research goals. Also, in the sections 3.1 – 3.5 (education block from the model), we presented our approach in which some activities carried more impact than others did. For example, the opportunity coast was high for hosting weekly open-door sessions when the attendance was low; however, open-door sessions proved very effective during the launch of Orion cluster in October 2015. Based on our observations we discontinued regular open-door sessions and replaced it with a more robust knowledge base, Slack, and one on one sessions when needed.

Overall, our activities throughout 2016 have resulted in significant growth in our research computing community with over 328 users as of March 2017 across Orion and DICE clusters (Figure 4). CDER cluster provides more than 700 users. In Figure 5, we provide the breakdown by percent of the HPC usage of GSU and XSEDE resources by month. Orion had 1.5 million CPU hours utilized in 2016, and XSEDE usage was at 2.7 million CPU hours (14.9 million XSEDE SUs) while maintaining an average 1.7 user-expansion factor [15].

A new management platform, ACoRE, featuring SLURM scheduler, integrates the three independent clusters: CDER, Orion, and DICE. Through ACoRE, we provide a similar environment to users as on national resources part of XSEDE. This creates a greater familiarity between local cluster(s) and national resources; as a result, simplifies user transition to national resources. To address the researchers' response of a need for greater technical training in our CI survey, we are launching a series of workshops that would focus on Introduction to Fundamentals in Data Science with a focus on Python and R technologies.

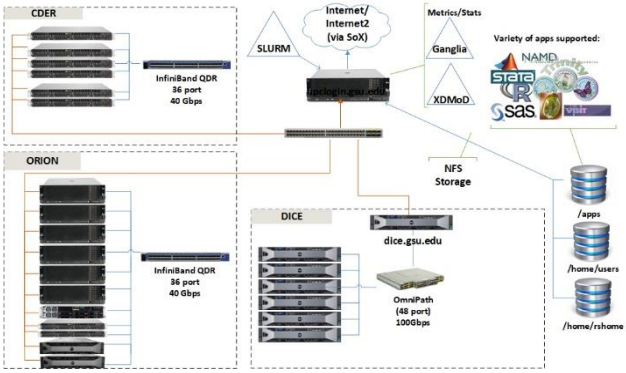

Figure 3. Advanced Computing for Research and Education (ACoRE) management platform for accessing Orion (360 core, 4.35 TB RAM), CDER (280 core, 960 GB RAM), and DICE (112 core, 896 GB RAM) clusters that are connected via a network supporting up to 10 Gbps.

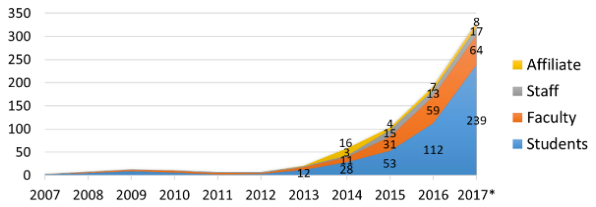

Figure 4. Research computing community across Orion and DICE clusters. This figure does not reflect more than 700 CDER users.

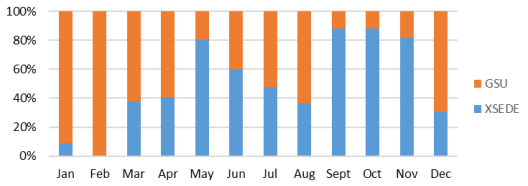

Figure 5. Total percent used of CPU hours from XSEDE and Georgia State [15-17].

5. CONCLUSION AND FUTURE WORK

We presented our workforce development model for training and preparing students for the Data Science field. Our model relied on integrating HPC resources within existing computational classes. Additionally, we created a series of workshops to supplement the classes and provide a venue for students and researchers representing *long tail of science* to learn the techniques in utilizing the resources for their research. In supporting our activities, we restructured our cyber-infrastructure by consolidating three clusters under one management platform – ACoRE, which enables users to take advantage of greater resources, reduces the burden for users to work with different schedulers, and reduces the burden for administrators to manage different application stacks and schedulers. This approach saves time on the administration of the clusters, and it enables facilitators to work with researchers on their research challenges. Using the SLURM scheduler across our resources enables a smoother transition from local resource to national resources for some of our more intermediate to advanced users as most national resources utilize SLURM.

6. ACKNOWLEDGMENTS

The National Science Foundation (NSF) under the grant CNS-1205650 supports CDER cluster that is led by PIs: Dr. Sushil K. Prasad, Dr. Anshul Gupta, Dr. Arnold Rosenberg, Dr. Alan Sussman, and Dr. Charles Weems . Georgia State University under the grant, 17-IST-107, supports DICE cluster, and under the grants 17-IST-104 and 18-IST-044 Scientific Computing Day. This work was supported by XSEDE Campus Champion Grants: GEO150002 and TRA130030. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by NSF grant number ACI-1053575. Authors would like to thank Dr. Sushil Prasad and Dr. Raj Sunderamman for approving the integration of CDER with ACoRE, and for their guidance and support in HPC curriculum development. Furthermore, authors would like to acknowledge Paul Anthony Bryan and Michael McDermott for the support they provided for CDER cluster prior to the integration into ACoRE. The authors would like to thank Brock Davis, Yuriy Lukinov, Michael Walters, and Dr. Kelly Stout for feedback that helped improve the manuscript. We would like to thank Michael Walters and Yuriy Lukinov for their support on Orion, DICE and CDER clusters.

7. REFERENCES

[1] Howe, H., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D.P., Kania, R., Schaeffer, M., Pierre, S.S., Twigger, S., White. O., Rhee, S.Y., Big data: The future of biocuration. Nature, 2008. 455(7209): p. 47-50

[2] Nelson, R., Collecting data, from sensors to systems. EE: Evaluation Engineering, 2014. 53(6): p.14-21.

[3] Kim, H.J., A. Pelaez, and E.R. Winston, Experiencing Big Data Analytics: Analyzing Social Media Data in Financial Sector as a Case Study. Proceedings for the Northeast Region Decision Sciences Institute (NEDSI), 2013: p. 62-69.

[4] Savage, N., Bioinformatics: Big data versus the big C. Nature, 2014. 509(7502): p. S66-S67.

[5] The White House. "Executive Order -- Creating a National Strategic Computing Initiative." July 29, 2015. Office of the Press Secretary. https://obamawhitehouse.archives.gov/the-press-office/2015/07/29/executive-order-creating-national-strategic-computing-initiative10.1145/2949550.2949584

[6] Advanced Cyber-Infrastructure for Research and Education Facilitators (ACI-REF). https://aciref.org/

[7] XSEDE Champus Champion Program. https://www.xsede.org/campus-champions

[8] Neeman, H., Bergstrom, A., Brunson, D., Ganote, C., Gray, Z., Guilfoos, B., Kalescky, R., Lemley, E., Moore, B. G., Ramadugu, S. K., Romanella, A., Rush, J., Sherman, A. H., Stengel, B., and Voss, D. 2016. The Advanced Cyberinfrastructure Research and Education Facilitators Virtual Residency: Toward a National Cyberinfrastructure Workforce. In *Proceedings of the XSEDE16 Conference on Diversity, Big Data, and Science at Scale* (XSEDE16). ACM, New York, NY, USA, , Article 57 , 8 pages. DOI: http://dx.doi.org/

[9] Prasad, S. K, Gupta, A., Kant, K., Lumsdaine, A., Padua, D., Robert, Y., Rosenberg, A., Sussman, A., Weems, C. "Literacy for All in Parallel and Distributed Computing: Guidelines for an Undergraduate Core Curriculum," CSI Journal of Computing, v.Vol 1.2, 2012.

[10] NSF/IEEE-TCPP Curriculum Initiative. https://grid.cs.gsu.edu/~tcpp/curriculum/?q=node/21183

[11] Rob Mitchum. "Unwinding the 'Long Tail' of Science." October 9, 2012. https://www.ci.uchicago.edu/blog/unwinding-long-tail-science

[12] Scientific Computing Day 2016 Symposium. September, 30 2016. Georgia State University. http://scd.gsu.edu

[13] Sarajlic, S., Edirisinghe, N., Lukinov, Y., Walters, M., Davis, B., and Faroux, G. 2016. Orion: Discovery Environment for HPC Research and Bridging XSEDE Resources. In *Proceedings of the XSEDE16 Conference on Diversity, Big Data, and Science at Scale* (XSEDE16). ACM, New York, NY, USA, Article 54, 5 pages. DOI: http://dx.doi.org/10.1145/2949550.2952770

[14] Sarajlic, S., Edirisinghe, N., Lukinov, Y., Walters, M., Davis, B., and Faroux, G (2016): HPC Strategic Model for Georgia State University: Discovery Environment for HPC Research and Bridging XSEDE Resources. Scientific Computing Day 2016. figshare. https://doi.org/10.6084/m9.figshare.3969375.v4

[15] Palmer, J. T., Gallo, S. M., Furlani, T. R., Jones, M. D., DeLeon, R. L., White, J. P., Simakov, N., Patra, A.K., Sperhac, J., Yearke, T., Rathsam, R., Innus, M., Cornelius, C. D., Browne, J.C., Barth, W.L., Evans, R.T., "Open XDMoD: A Tool for the Comprehensive Management of High-Performance Computing Resources",Computing in Science & Engineering 17.4 (2015):52-62, 2015 10.1109/MCSE.2015.68

[16] Open XDMoD for Georgia State University. http://xdmod.rs.gsu.edu

[17] XDMoD. https://xdmod.ccr.buffalo.edu/