Shantenu Jha*[1] and Heather Lynch and Allen Pope and Jane Wyngaard
* *Electrical and Computer Engineering, Rutgers University*
+ *Molecular Physiology and Biological Physics and Biomedical Engineering, University of Virginia*

Abstract

The NSF funded Polar Computing Research Coordination Network NSF 1542110 (http://polar-computing.org) was tasked with analyzing opportunities and barriers in the uptake of high-performance & distributed computing for polar science. Specifically the charge of the RCN was to: (i) Identify opportunities for, and barriers to, greater uptake of High-Performance and Distributed Cyberinfrastructure by polar sciences, (ii) Ensure that plans and designs for new and existing NSF-funded cyberinfrastructure efforts are cognizant of the needs of the Polar Science community, and (iii) Understand how best to educate a new generation of polar scientists in the skills needed to realize the opportunities and potential of HPDC. This community paper analyzes the state-of-play along these three dimensions.

Introduction

Climate change is having and will have a vast impact on the polar regions including sea ice retreat, mass loss from the Greenland and Antarctic ice sheets, ecosystem disruption and change, and retreat of mountain glaciers, not to mention additional impacts on precipitation regimes, extreme events, and other key environmental processes. A deeper understanding of these changes requires enormous quantities of diverse observational data (e.g. from satellite and airborne imagery to automated field observations), the ability to integrate these streams of data into detailed computer models, and sophisticated means of analysis and visualizing emergent patterns. Efficient analysis of the increasingly large and complex volumes of data produced by the Intergovernmental Panel on Climate Change (IPCC) studies and others require greater sophistication of computing, visualization, and data management. These are emblematic of the challenges required to solve these and other key problems in the polar sciences, for which the efficient utilization of high-performance and distributed computing has become increasingly critical to continued scientific advancement.

In addition to scaling and increasing the sophistication of existing approaches, there are unprecedented new opportunities and fundamentally different solutions emerging. It is no exaggeration that the onset of Machine Learning and other Artificial Intelligence methods have the potential to radically change the state-of-practice of Polar Sciences, as well as many observational data rich domain sciences. These new opportunities in turn require the full utilization of existing and emerging high-performance computing capabilities.

Prior to the inception of a Research Coordination Network (RCN) on Polar Computing, at least two workshops had demonstrated the need for advanced cyberinfrastructure [28] and high-performance and distributed computing (HPDC) to solve different polar scientific problems [Ref 2, 32]. The penetration of HPDC in polar science has been low relative to other domains and has not increased over time in proportion to the opportunities. For example, there were no polar science projects in the XRAC allocation portfolio at the time the RCN was commissioned. In response, this National Science Foundation funded

[1] shantenu.jha@rutgers.edu

Research Coordination Network (RCN) on Polar Computing (led by Jha, Lynch, Wyngaard, Nabrzyski, & Yarmey/Pope http://polar-computing.org) was tasked with identifying opportunities, structural barriers, and proposals for how to overcome these barriers to bridge the current gap between the polar science and HPDC communities.

The RCN organized several stand alone workshops and hackathons, organized special sessions at relevant workshops/conferences (e.g., XSEDE 2016, SCAR 2016, Polar 2018) and participated in other community activities (e.g., other EarthCube and NSF organized meetings). It has identified several bottlenecks that currently limit the uptake of HPDC in the polar sciences. Whereas there are âĂIJbleeding edgeâĂİ aspects of HPDC that are technically demanding and will always remain difficult to use, there remains a perception that initial access and basic utilization remains a fundamental barrier due to inadequate training and training opportunities for domain scientists. To this end, the RCN has undertaken several community-wide education and training activities, some quite general (e.g., using a shell environment) and some more narrowly targeted (e.g., HPC tools for high-resolution imagery), that in aggregate provide a solid foundation for future progress.

As the Arctic DEM and Reference Elevation Model of Antarctica (REMA) projects [27] demonstrate, there are clearly some polar science projects at the vanguard of using HPDC. These and other projects [Ref 17] demonstrate the potential impact, but such success are just emerging from the community. Beyond large scale projects like the ArcticDEM, polar science is also characterized by many, more highly focused projects often referred to as the âĂIJlong tail of scienceâĂİ. These projects may be narrower in scope they can require vast quantities of HPC, storage and data. Whereas the RCN focused on ensuring that HPDC infrastructure works for projects of all sizes and scope, it primarily focused on identifying how to increase HPDC penetration across a larger swath of the polar science community.

This report "The Landscape of High performance and distributed computing in polar science" aims to:

i. identify barriers to greater uptake of High-Performance and Distributed Cyberinfrastructure by polar sciences,
ii. highlight opportunities for HPDC to contribute to advances in data analysis & management, simulation, and modeling of complex processes.
iii. create an awareness for cyberinfrastructure efforts to be cognizant of the needs of the Polar Science community.
iv. understand how best to educate a new generation of polar scientists with skills needed to realize the opportunities and potential of HPDC.
v. advocate for continued partnerships between polar and computer sciences, and
vi. quantify the âĂIJbase levelâĂİ of HPC required to produce the core products that are now required for researchers to meet their funded science goals.

Importance of Community Oriented Research Cyberinfrastructure

The term cyberinfrastructure "CI" is defined as a research environment, accessible over the internet, that supports all data processing steps from data acquisition, storage, management, integration, mining and visualization [9]. In addition to the software and hardware components, cyberinfrastructure also includes the personnel required to create and operate these resources and specialized advisors who can assist researchers in using these resources. In short, effective CI efficiently connects labs, data, computers, and researchers to facilitate scientific discovery. Perceived advantages of organized and balanced cyberinfrastructure include:

i. maximizing the shared benefits between the scientific organizations,
ii. eliminating or minimizing the redundant investments in these fields and

iii. avoiding the current barriers that can slow down the progress of research.

Although organized and balanced CI is necessary, judging by examples from Molecular Science and High-Energy Physics, domain specific community CI efforts seem to be necessary to build upon the general CI and to address domain specific requirements. The Polar Computing RCN represented the polar community's first steps towards an organized CI that meets the requirements of the long-tail of the polar science community. The Polar Computing RCN also identified the primary and critical reasons behind a gap in the skill set and usability (simplicity) of tools.

Existing Projects in the Polar Sciences that make use of advanced and scalable CI

We briefly mention some significant projects that are working to bring scalable Cyberinfrastructure to the Polar Science community. Some of these are sophisticated and scalable consumers of HPDC (ArcticDEM and REMA) with requirements long defined by the science community (NRC, 2007, 2017), others are smaller projects that use local computing resources. Taken together, they build on and represent several significant investments in centers responsible for data repositories and computation.

ArcticDEM: ArcticDEM is an NGA-NSF public-private initiative to automatedly produce a high-resolution, high quality, time-dependant, digital surface model (DSM) of the Arctic using optical stereo imagery, high-performance computing, and open source photogrammetry software. ArcticDEM encompasses all land area north of 60ÂřN. In addition, coverage includes all territory of Greenland, the State of Alaska in entirety, and the Kamchatka Peninsula of the Russian Federation. ArcticDEM data is constructed from in-track and cross-track high-resolution (0.5 meter) imagery acquired by the DigitalGlobe constellation of optical imaging satellites and licensed through the NGA NextView contract. The majority of ArcticDEM data was generated from the panchromatic bands of the WorldView-1, WorldView-2, and WorldView-3 satellites.

REMA: The Reference Elevation Model of Antarctica (REMA) is a high resolution, time-stamped Digital Surface Model (DSM) of Antarctica initially at an 8-meter spatial resolution. It provides the first, high resolution (8-meter) terrain map of nearly the entire continent. REMA is constructed from millions of individual stereoscopic pairs of submeter imagery; each individual DEM was vertically registered to satellite altimetry measurements from Cryosat-2 and ICESat, resulting in absolute uncertainties 1m over most of the continent , and relative uncertainties of decimeters. Since each REMA grid point has a timestamp, any past or future point observation of elevation provides a measurement of elevation change. REMA may provide corrections for a wide range of remote sensing processing activities, such image orthorectification and interferometry, and provide constraints for geodynamic and ice flow modeling, mapping of grounding lines, and surface processes. REMA also provides a powerful new resource for field logistics planning.

ICEBERG: Imagery CI and Extensible Building-Blocks to Enhance Research in the Geosciences: The ICEBERG Project [15] is building the cyberinfrastructure required to do imagery-enabled science at very large (e.g., continental) scales. Use cases to be developed include a Southern Ocean pack ice seal survey using deep learning and a pan-Antarctic land cover classification product. One of the emerging themes of the ICEBERG project is that machine learning and algorithms drawn from computer vision play a key, and perhaps underappreciated role in the future of imagery and big data-enabled geosciences, and they require specific cyberinfrastructure (GPUs, for example) that are less widely available.

Engaging the Greenland Ice Sheet-Ocean (GRISO) Science Network: The goal of this project [3] is to enhance multi-disciplinary activities and collaboration through the establishment of the GRISO (Greenland Ice Sheet-Ocean) Science Network using scientists and cyberinfrastructure experts.

Polar Globe Power Up Polar Cyberinfrastructure using M-cube Visualization to support polar Climate Data: The Polar Globe [21] develops and implements visualization techniques that will serve as an Arctic research platform and to support polar climate studies.

PolarGrid - Cyberinfrastructure for Polar Science: The Polar Grid project at Indiana University was designed to help scientists better understand and examine the state of polar ice sheets. The PolarGrid NSF Project supported advanced computational methods to process a large amount of the collected data from the polar regions. PolarGrid used the open data access standard technology to make the collected data processing, simulating outside the PolarGrid and for the majority of the polar science community.

Center for Remote Sensing of Ice Sheets (CReSIS): In 2005, NSF established CReSIS with the goal of enhancing the quantity, quality, and use of radar data collected over the Greenland and Antarctic ice sheets. Since 2010, Indiana University and the Center for Remote Sensing of Ice Sheets (CReSIS) at the University of Kansas have worked with NASA's Operation IceBridge to collect data about polar ice caps. IU provides IT support for the missions and assists in processing the enormous amounts of data the mission generates, helping improve the models of the physical interactions of glaciers, sea ice, and ice sheets for scientists to study. IU has created on-plane storage and computational systems used by scientists during flights to monitor activities. What once took researchers 18+ hours to view first glance data products now takes less than 10 seconds thus allowing researchers on those flights to make adjustments in real time. CReSIS REU Site was established with ideas of supporting activities in Polar science, education, diversity and outreach and will train the next generation of scientists, engineers and educators interested in polar science and cyberinfrastructure.

National Snow and Ice Data Center (NSIDC): NSIDC supports polar and cryospheric research. NSIDC manages and distributes scientific data, creates tools for data access, supports data users, performs scientific research, and educates the public about the cryosphere.

Polar Geospatial Center (PGC): The PGC at the University of Minnesota provides geospatial support, mapping, and GIS/remote sensing solutions to researchers and logistics groups in the polar science community. PGS was integral to development and leads support of both ArcticDEM and REMA described above, and has been an active participant in many related RCN efforts.

Barriers to greater HPDC uptake

Polar Sciences has a mix of black-box HPC users and scientists who need to develop methods and algorithms for use on HPC resources. For both types of polar researchers HPDC resources and capabilities should be delivered as a seamless, scalable service. However, both user types face barriers arising from an intrinsic and often times unavoidable complexity in using HPDC resources. Some, but not all the complexity of using HPDC resources can be addressed by adequate experience and exposure to software best practices, as provided in training such as software carpentry. Further, many researchers are simply unaware of computing resources available to them, whereas others may lack the appropriate training required to take advantage of HPDC resources they may already have access to. Put together, these reinforce the importance of education and training.

Education and Training

Despite these data- and compute-intensive scientific needs, polar science (along with other disciplines) is poorly represented in the use of HPDC resources. Through informal community engagement, a Polar-HPDC workshop in 2014 [2], and a 2016 hackathon, the Polar Computing RCN identified two primary reasons for this gap:

1. a lack of community awareness of available compute resources, and
2. a lack of appropriate training for scientists interested in using HPDC for science applications.

Over the last three years, during which time five major RCN-related events have taken place, the Polar RCN has investigated community needs via surveys to the community, workshops, and training events designed for both researchers who were actively seeking to use HPDC and those who not aware of the possible value of such to their work.

In the first year, we hosted a week-long hackathon co-located with the XSEDE annual meeting in Miami, ran a short side meeting at the Scientific Committee for Antarctic Research Open Science Conference, and led a 2-day workshop on imagery-enabled science at the PGC (see above). In the second, we ran a stand-alone week-long event at Stony Brook University's Institute for Advanced Computational Science. The RCN's 3rd major training event was co-located with the Polar 2018 Open Science Conference in Davos, Switzerland. Through our event experiences and with increasing longer-term community engagement we tuned the execution of these events but in all cases found the community highly appreciative and urgently requesting more such events.

Aspects of our approach that we found highly successful and which we retained through all three training events were:

1. to include examples of how HPDC might be useful to polar domain science in our event advertising,
2. survey attendees prior to the event to gauge their breadth of HPDC relevant skills, and
3. to require attendees propose a particular research question to which they wanted to apply HPDC resources.

The two most significant changes made following the first training event were to:

1. to increase the amount of direct classroom training provided (by inverting the time ratios of hacking to training), and
2. to bring teams of researchers to a hackathon or training event rather than form teams from multiple institutions.

This last change was particularly beneficial because it minimized the amount of time spent "learning how to work as a group", and it provided more opportunities for teams to continue working on challenges after the event was finished.

In increasing the amount of formal training at our events, we found the materials available through the Software Carpentry (SWC) Foundation invaluable. HPC Carpentry was unfortunately not available in its current form at the time, however, the SWC materials on bash, git, and python were particularly useful and provided the right balance of demonstration and hand-on practice. We found the lack of polar-specific examples to have at most a minor impact on learning of the basic materials, and were able to supplement the pre-existing lessons with examples of how polar scientists were using HPDC in their work. While increasing the amount of time spent in formal classroom instruction requires finding sufficient volunteer

tutors, SWC has done an excellent job training new instructors and the SWC instructors at our final event were both polar researchers. In our first event we had attempted to include computer science experts in the hackathon, under the assumption that offering the potential for new collaborations might be an attractive draw. However, this was more difficult than anticipated, and the difficulty of engaging computer scientists in this work (as opposed to training domain scientists in computation) remains a stubborn challenge for which we have no easy remedies.

The challenge in the above led in part to the second primary alteration to our approach. While from the beginning we had request attendees come with a project in mind, and bring collaborators for such, as we increased the formal training time it became all the more important that they bring a team of skilled collaborators, many but not all of which included research computing or HPDC experts.

The key education and training challenges we have identified are:

1. Pace of technological change: HPDC tools and available resources are changing too fast for a non-domain-specialist to keep up to date.
2. Momentum: There is significant momentum behind using known and trusted methods in an environment that does not provide for high risk non-domain exploration.
3. Limited Time: Across the spectrum of undergraduate through experienced PI, practicing and training domain scientists do not have the âĂŸbandwidthâĂŹ to accommodate additional core skills training.
4. Inappropriate material: Most HPDC infrastructure resources have training materials but these are overwhelmingly targeted at a computer expert audience and therefore assume a significant amount of knowledge and even language terms that domain scientists are unlikely to have or understand.
5. Limited HPDC Support Staff on Campuses: Only campuses with large scale HPDC infrastructure appear to have sufficient on-site staff with the time to train and work with domain scientists.
6. High Barriers: For a non-expert, the final step of actually porting research onto an HPDC resource, even after extensive abstract training, is often simply still a barrier too high to be crossed without intensive support.

Until now there has not been a collaborative effort to build the proper materials and support structure to address the above challenges. The first vision of HPC carpentry is currently being tested and refined for release. This will go a long way to supporting the needed short-form, intensive, non-expert targeted type of materials. Institutional or NSF support for running training events and maintaining and improving HPC carpentry materials – including tuning them to domain specific needs – would be beneficial. Based on our experiences, we would recommend that training events be paired with hackathon-type tutored time allocations wherein attendees are able to take the final, most daunting and difficult, steps to running and developing their work on an HPDC infrastructure with guidance and support from trained HPDC experts.

Commercially Licensed Software

One of the big barriers to the uptake of HPDC we identified over the course of our RCN is the community's reliance on expensive licensed software that is not easily deployed on HPC systems. This is often a particular barrier when multiple analyses are required to run in parallel, because licenses may restrict deployment to only a small number of cores. As a result, domain scientists may be faced with re-writing these workflows using open-source tools. This dynamic has been playing out, for example, with the replacement of ESRI's ArcGIS with the open-source QGIS in many geospatial workflows. One of the major successes in

polar cyberinfrastructure is Quantarctica, a free GIS package and data collection built on QGIS and including a large and growing number of community-developed geospatial datasets packaged for interoperability; development of regional Arctic QGIS-based GIS packages would be a boon to the community.

A Predictable Supply of HPC for Large-scale Community Projects

The polar science community's dependance on data produced on NSF funded HPC became clear after the release of ArcticDEM and REMA. PI's quickly applied the DSMs to science that occurs on the surface of the earth, including geomorphology, the carbon cycle, population dynamics, and ice mass balance. The production of DSMs is computationally demanding, and is currently centralized activity for which the PGC is responsible. The production of DSMs presents unique policy and allocation considerations. Currently as the central data provider, the PGC must compete for HPC allocations to produce DSMs; however the entire community requires the procesed end-products. It is imperative that predictable and sustainable access to HPC resources be arranged to ensure the uninterrupted capability to generate DSM and other products upon which the community critically depends.

Opportunities

An analysis of other scientific domains suggests as the penetration of HPDC (and computing as a whole) increases, the impact of HPDC changes. In initial stages, scientists do simple things faster, maybe larger and possibly some times even better. As the level of sophistication with HPDC increases, the use of HPDC evolves from solving 'old' problems faster, to solving 'old' problems better (improving estimates of uncertainty, for example), to actually solving new problems that were not previously feasible. This HPDC transition has been borne out in many other communities, such as molecular sciences and engineering, and there is no reason to suspect that such a transition is not also possible in polar sciences.

Simulation and modeling of physical phenomena have historically been major consumers and drivers of high-performance distributed computing resources. Although significant advances have been made in both the cyberinfrastructure (software, data analysis etc.) for simulation & modeling, and the resulting science, there remain ongoing challenges in transitioning to the next generation of HPC platforms and architectures, as well as in ensuring the software capabilities stay current. The challenges faced and opportunities for the Polar simulation and modeling community are mostly in line with and shared by the broader simulation and modeling community. With no intention of undermining the opportunity arising from advances in simulation and modeling, we highlight a couple of new opportunities that have high potential return-on-investment for the polar science community, both for individual researchers and for federal investment for polar cyberinfrastructure.

Cloud Computing

Cloud computing presents significant benefits for scientists, including those working on polar applications. Unlike XSEDE and other federal HPDC resources, cloud computing is portable from one institution to the next and requires only a credit card, rather than a formal proposal, for access - two benefits especially important to early-career scientists just getting started using HPDC resources. Because cloud service providers are vying for business, the user interfaces are more approachable, the documentation is often less technical, and the barriers to entry and significantly lower for new users. As users gain in experience and sophistication, the cost of running analyses in the cloud is often an incentive to move to XSEDE or local resources, by which time the user has a greater understanding of their compute needs and is in a better position to write a successful proposal for computing time. In this way, cloud computing can be a gateway to HPDC resources for new users and, over the long haul, provides an opportunity to maintain continuous support for computing pipelines that might continue development long after the initial grant (or access to XSEDE) has concluded.

Machine Learning & Computer Vision

Machine learning (ML) is perhaps one of the most exciting new areas for polar science, in no small part because of the rapid proliferation of high resolution imagery. Our own consideration of ML opportunities are focused on the intersection of machine learning and computer vision, because this was the issue that was raised most frequently in our interactions with polar scientists. Unlike many of the challenges and opportunities uncovered by the RCN, this issue is in some ways uniquely pressing for polar science because of the greater access to commercial imagery in the polar regions compared to other areas. ML presents almost unlimited opportunities for polar science that the community has only begun to tap. Dozens if not hundreds of studies have demonstrated the utility of satellite imagery for polar science, though to date the majority of these applications have involved "traditional" pixel-level classification methods (e.g., clustering, spectral angle mapping, spectral mixture analysis, supervised maximum likelihood classification) to assign classifications (sea ice, land, vegetation, penguin guano, etc.) to each image pixel; some work has also been done with object-based image analysis which aggregates groups of pixels. Until recently, medium resolution sensors such as Landsat and MODIS represented the workhorses of polar science. More recently, interest has expanded to include a suite of high-resolution (submeter) commercial satellite systems (WorldView, Quickbird, etc.). Not only are these higher resolution systems being used to study 'old' problems in greater detail, but new applications have been developed. One such application is in biology, since sub-meter imagery is now making it possible to track both Antarctic wildlife (penguins, seals) and map vegetation at spatial scales not previously within reach. While the spatial resolution of this imagery presents unprecedented opportunities, it also presents two major challenges:

1. It is not feasible to manually annotate or interpret imagery at large spatial scales.
2. The automated methods for doing so require not only new developments in machine learning but a significant investment in HPDC.

This requires not only new systems for uniting imagery and HPDC but also a greater emphasis on training polar scientists to use HPDC (Section 3). A new NSF EarthCube project ICEBERG is dedicated to building cyberinfrastructure to solve the first of these challenges, though longer term investment in collaborations between the machine learning community and the polar community will be needed if the polar community is to fully reap the benefits of its commercial imagery assets.

Recommendations

1. Lowering the barrier to access to National CI: Lower the barrier to accessing National CI through targeted appropriate education and training. Support use of commercial, cloud based HPC resources as a gateway to institutional and federal HPC resources.

2. Representing the needs of Earth Science in the design and analysis of National CI: The design and implementation of cyberinfrastructure resources (e.g., XSEDE) have been dominated by the traditional needs of the simulation and modeling community. As alluded to several times in this report, Polar Sciences is representative of a number of domains in the Earth Sciences that have become observational data driven. Specifically, many of the challenges and lessons from polar imagery are applicable to other subdomains in the Earth Sciences. Thus, not only should the cyberinfrastructure needs for the Polar Science community be understood, but they must be articulated and addressed in the context of national and international cyberinfrastructure efforts. In other words, existing and future (NSF funded) cyberinfrastructure efforts must be aligned and integrated with the requirements of Polar Science community.

3. **Education and Training** Lack of education and training remains a significant barrier to HPDC uptake. It presents as both a lack of awareness and as a skills and jargon barrier. This lack requires a response that is targeted at researchers at every career stage (undergrad through experienced PI) and which is therefore accommodating of such (time, knowledge, and need limitations). Based on both the experiences of the RCN and other documented research, we recommend The Carpentries materials be used as an immediately available, sustainable, and level appropriate source of training materials, that are effective when used in short, highly tutored, team based, and research project focused training events. These events were found to be efficient means of moving teams of researchers from limited laptop and local server computational resources to appropriate HPDC resources. Lack of resources to run such events including necessary tutor time, and lack of awareness of the need and mismatch of existing available materials by National CI facilities is limiting the availability of such events and better training resources.

4. **Sustained Investment in Collaboration:** Cyberinfrastructure for polar science is inherently and unavoidably interdisciplinary. This requires domain scientists and cyberinfrastructure experts to come together, understand a common language, and work together to meet community needs. The key feature of this kind of research is that it requires teams, often fairly large teams, to effectively translate scientific questions and technical needs into functioning cyberinfrastructure that can be understood and readily adopted by the community. It is often challenging to assemble such teams and operate efficiently under the auspices of a short (2-3 year) award, and it is often difficult to recruit research software engineers to work on such short term projects. Many of the most successful teams are quasi-permanent assemblages of domain scientists, computer scientists, research programmers and cyberinfrastructure experts, often working at institutes or national labs that can support personnel over extended periods of time beyond a single award. Sustained innovation in polar cyberinfrastructure is likely going to require opportunities for funding that extend beyond a typical grant cycle, can support community management, or funding of institutions and national facilities where bridge funding can be used to cover gaps between awards.

5. **Predictable and Sustainable Computing Allocation** A large number of funded science goals are now dependant on products such as ArcticDEM, REMA and the Airborne RADAR processing. resources. The production of such community products is computationally demanding and depends upon a predictable and sustainable HPC computing resource allocation(s). This requires a new perspective on community computational capabilities and allocation to ensure adequate and uninterrupted capability to generate DSM and other products upon which the community critically depends.

Acknowledgements:

References