2018 Symposium of the Center for Network and Storage Enabled Collaborative Computational Science

Monday 15 October 2018 - Tuesday 16 October 2018

University of Michigan

Book of Abstracts

This is the current set of abstracts for the CNSECCS Symposium on October 15-16, 2018.

Contents

Science Use-Cases / **6**

SAFRN: Statistics computed without sharing private data

Authors: George Alter[1]; Rafail Ostrovsky[2]; Steve Lu[3]; Brett Hemenway Falk[4]

[1] *University of Michigan*

[2] *UCLA*

[3] *Stealth Software Technologies*

[4] *University of Pennsylvania*

Corresponding Author: altergc@umich.edu

Although government agencies, health providers, and organizations of all kinds collect and store data at an ever-increasing rate, much of the most valuable data is unavailable for research and policy due to the need to protect privacy. Extracting relevant information is especially difficult when research requires data residing in multiple locations with separate privacy controls. Stealth Software Technologies has partnered with ICPSR to implement a privacy-preserving statistical system that enables data analysis without requiring data owners to reveal any information about individuals to the analyst or to each other. Secure Multiparty Computation (MPC) is a framework that enables multiple data-holders to compute any joint function of their private inputs, while maintaining the privacy of each input. SAFRN will compute standard statistical analyses (crosstabulation, difference of means, regression, logistic regression) while provably maintaining the privacy of each participant's confidential information.

Related Projects / **17**

Collaborative Workflow-Driven Science in a Rapidly Evolving Cyberinfrastructure Ecosystem

Corresponding Author: altintas@sdsc.edu

Scientific workflows are powerful tools for computational data scientists to perform scalable experiments, often composed of complex tasks and algorithms distributed on a potentially heterogeneous set of resources. Existing cyberinfrastructure provides powerful components that can be utilized as building blocks within workflows to translate the newest advances into impactful repeatable solutions that can execute at scale. However, any workflow development activity today depends on the effective collaboration and communication of a multi-disciplinary data science team, not only with humans but also with analytical systems and infrastructure. Dynamic, predictable and programmable interfaces to systems and scalable infrastructure is key to building effective systems that can bridge the exploratory and scalable activities in the scientific process. This talk will focus on our recent work on the development of methodologies and tools for effective workflow driven collaborations, namely the PPoDS methodology and the SmartFlows family of tools for the practice and smart utilization of workflows.

Lightning Talks / **13**

You have data, we have questions!

Authors: Susan Borda[1]; Scott Witmer[1]

[1] *University of Michigan - Library*

Corresponding Author: sborda@umich.edu

Researchers put considerable time and effort into research, and the resulting data is a significant scholarly product. As with papers/articles/presentations, data should be treated like a first-class scholarly/research product that requires additional considerations. Publishers and funding agencies are requiring researchers to share supporting data, for example. Repository managers, data curators, and archivists would like to assist researchers in meeting such data requirements. However, many questions arise in the minds of those who would like to assist researchers, especially when it comes to more complicated data such as variety at the center of this symposium, large, collaborative, model or simulation driven computational data. This lightning talk will address important questions for researchers to consider as they deposit their research data.

Here is a sampling of the issues researchers should think about:

Reproducibility goals:
Is there a further requirement for reproducibility or replicability? If so, is the dataset complete?

Preservation goals:
Is it worth it to keep simulation data long-term, more than ten years? If not, should anything remain, what would be of use to future researchers? Does the evolution and development of better/faster simulation software and computing technology over time make the preservation of older simulation data redundant? As it becomes easier to replicate simulations and rerun them as needed, does the actual data output need to be preserved, or just the inputs/parameters? Is there some simulation output data worth keeping? If so, what are the criteria?

Re-use goals:
Should the raw data be shared or only the "final,"analyzed data that directly support the figures in the paper? Could the raw data be useful to someone else in the same discipline or another one? What role do repositories play in the potential reuse of data? If researchers want to reuse someone else's data, are they more likely to access it through a repository or contact the researcher directly? Libraries keep talking about data re-use as an important driver for data sharing, but is it?

Answering these questions supports both the immediate goals of the dataset as well as anticipating what will be necessary to preserve and possibly re-use the data. Depending on the real needs of the researcher and dataset, the files and information included with the deposit may need to change. Data retention timelines may differ as well, depending on the method used to generate it.

Science Use-Cases / **4**

Big data challenges in understanding and modeling multiphase flows

Author: Jesse Capecelatro[1]

[1] *University of Michigan*

Corresponding Author: jcaps@umich.edu

Multiphase flows are ubiquitous in both engineering applications and the environment. Within the energy sector, such flows play fundamental roles in chemical transformation reactors, spray combustors, and slurry pipelines, to name just a few examples. Understanding and predicting such flows is key to ensuring optimal performance and improving design. Environmental processes such as gravity currents, hurricanes, debris flows, and atmospheric dispersion of pollutant particles also represent multiphase flows with great societal importance. Improving our understanding of these flows is critical to develop strategies to control them and mitigate their negative effects. The key challenge is attributed to the wide range of length- and time-scales inherent to these flows, which typically vary by many orders of magnitude. The advent of modern computing has led to significant progress in numerical modeling of multiphase flows, yet simulation capabilities at practical engineering and environmental scales still remain out of reach. In this talk, I will highlight the current state-of-the-art in numerical simulations of turbulent multiphase flows and highlight the key limitations and challenges. I will argue that model development and new insights into these complex

flows are hindered by the availability of diverse large-scale datasets. Fast and open-source access to turbulent multiphase flow data would enable the scientific community to probe flow physics, test new models, and quantify uncertainty across different modeling approaches. The Johns Hopkins Turbulence Database is widely considered the 'go to' source for large-scale turbulence data. Meanwhile, an open-source database of canonical multiphase flows does not exist. The talk will conclude with a pithy description of what such a database might look like, and its potential impact.

Complementary Technology Solutions / **15**

Reproducible computational workflows with signac

Author: Bradley Dice[1]

Co-authors: Carl S. Adorf [1]; Vyas Ramasubramani [1]; Sharon C. Glotzer [1]

[1] *University of Michigan*

Corresponding Author: bdice@umich.edu

The open-source Python framework *signac* is designed to manage data sets and perform operations on the data in an efficient, reproducible, and collaborative way. The framework is particularly well-suited for data-driven exploration of file-based, dynamic and heterogeneous data spaces. In contrast to many databases and task executors, *signac*'s serverless data management and *signac-flow*'s portable workflow model ensure that workflows are just as easily executed on laptops as in high-performance computing environments. The *signac* approach not only increases research efficiency, it also improves reproducibility and lowers barriers for data sharing by transparently enabling the robust tracking, selection, and searching of data by its metadata. Collaboration on *signac* data spaces is as simple as using any shared network file system. In the last year, several features have been added to improve searching, synchronizing, importing, and exporting data.

Related Projects / **7**

Michigan Institute of Data Science: Computational Challenges and Research Opportunities

Author: Ivo Dinov[1]

[1] *University of Michigan*

Corresponding Author: dinov@umich.edu

Presenter: Ivo D. Dinov
In this talk, I will present the Michigan Institute of Data Science (MIDAS), a trans-collegiate Institute at the University of Michigan. I will start by describing the multidisciplinary activities in data science at the University of Michigan. Then I will cover some of scientific pursuits (development of concepts, methods, and technology) for data collection, management, analysis, and interpretation as well as their innovative use to address important problems in science, engineering, business, and other areas. We will end with an open-ended discussion of educational challenges, research opportunities and infrastructure demands in data science.

Science Use-Cases / **14**

Advanced Computational Modeling Tools for Patient-specific Hemodynamic Analysis

Author: C. Alberto Figueroa[1]

[1] *University of Michigan*

Corresponding Author: figueroc@med.umich.edu

Advances in numerical methods and three-dimensional imaging techniques have enabled the quantification of cardiovascular mechanics in subject-specific anatomic and physiologic models. Research efforts have been focused mainly on three areas: i) pathogenesis of vascular disease, ii) development of medical devices, and iii) virtual surgical planning.
However, despite great initial promise, the actual use of patient-specific computer modelling in the clinic has been very limited. Clinical diagnosis still relies on traditional methods based on imaging and invasive measurements. The same invasive trial-and-error paradigm is often seen in vascular disease research, where animal models are used profusely to quantify simple metrics that could perhaps be evaluated via non-invasive computer modelling techniques. Lastly, medical device manufacturers rely mostly on in-vitro models to investigate the anatomic variations, arterial deformations, and biomechanical forces needed for the design of medical devices.

Our laboratory has been developing an integrated image-based computer modelling framework for subject-specific cardiovascular simulation CRIMSON (www.crimson.software) that can successfully bridge the gap between the research world and the clinic.

In this talk, we will provide an overview of the most novel features for the software, specifically the functions for parameter estimation and simulation of transitional stages, and highlight a series of future developments for the project.

Lightning Talks / 12

Campus Integration projects: Building the base of a national cyberinfrastructure

Author: Richard Knepper[1]

[1] *Cornell University*

Corresponding Author: rich.knepper@cornell.edu

This lightning talk discusses two ways for campuses to extend their cyberinfrastructure capabilities and broaden offerings for researchers: toolkits from the XSEDE Cyberinfrastructure Resource Integration team, and the Campus Compute Cooperative group.

Campus resources, local clusters and storage, networking, and consulting, are the basic foundation for computational resources. Campus IT staff encounter challenges with implementing, maintaining, and extending these resources. Organizations such as XSEDE provide means for disseminating best practices, expertise, and configurations from national resources to campuses. The integration activities of XSEDE Cyberinfrastructure Integration group help campus IT professionals to implement and extend their local cyberinfrastructure in ways that allow researchers to leverage beyond the local to the national.

Likewise, the Campus Compute Co-Operative project is in the process of building an exchange in which members can use resources at other institutions, either through exchange of like resources or through payments. The cooperative allows a campus to volunteer resources that may be in excess supply in exchange for resources at other institutions that may have particular capabilities that are required. The cooperative allows its members to make use of a broader range of resources without forcing each campus to acquire all types of resources that may be in demand, extending the range of capabilities for members.

Related Projects / **16**

The OSiRIS Project

Corresponding Author: shawn.mckee@cern.ch

The NSF-funded OSiRIS project (http://www.osris.org), which is creating a multi-institutional storage infrastructure based upon Ceph and SDN is entering its fourth year. We will describe the project, its science domain users and use cases, and the technical and non-technical challenges the project has faced. We will conclude by outlining the plans for the remaining two years of the project and its longer term prospects.

Welcome and Overview / **18**

Symposium Logistics

Corresponding Author: shawn.mckee@cern.ch

Information about logistics for the Symposium

Welcome and Overview / **20**

Symposium Overview

Corresponding Author: shawn.mckee@cern.ch

Lightning Talks / **5**

Challenges of automated data transfer with ePHI

Author: Deb McCaffrey[1]

[1] *University of Michigan medical school*

Corresponding Author: debmccaf@med.umich.edu

Many researchers are interested in lab test results from external vendors beyond what is normally reported in the clinic. They have a need to retrieve the data from vendors, deidentify it, and send the deidentified data to the final analysis location. Automating this process is convenient for the research team and also helps protect the patient data further. However, this provides unique challenges for the transfers.

Welcome and Overview / **19**

Symposium Welcome Address

Corresponding Author: emichiel@umich.edu

Complementary Technology Solutions / **10**

Distributed Data in Astronomy: A Heterogeneous Network

Author: Christopher Miller[1]

[1] *University of Michigan*

Corresponding Author: christoq@umich.edu

Astronomy is a data intensive science that utilizes remote sensing. Astronomical data has always been heterogeneous in terms of its distribution, accessibility and characterization. Today, the data have become so large that one is required to conduct the analysis at the location of the data, creating new challenges for data security. I will discuss different solutions for how astronomy is addressing these issues, including modern big data machinery such as Johns Hopkins "SciServer" and the National Optical Astronomy Observatory's "DataLab".

Lightning Talks / **1**

Modern Multi-tier Storage in Campus Environments

Author: Brock Palen[None]

Corresponding Author: brockp@umich.edu

Data volumes and data variety on campus has grown drastically in recent years. ARC-TS, the research computing provider at the University of Michigan embarked on a project to provide mult-tier storage environment for our research community. The goals of this project was to meet several needs:

- Address the needs of both open-science and restricted data
- Apply performance levels and costs that match the data domain
- Support matching research life cycle, tracking the value of the data over time

This short talk will address how we addressed this need and our experiences and challenges along the way. We will also address what we think our next steps are to make data more accessible and portable between local and remote environments.

Funding Agency Perspectives / **3**

Supporting Advanced Research Cyberinfrastructure

Authors: Stefan Robila[1]; Amy Walton[1]

[1] *National Science Foundation*

Corresponding Author: srobila@nsf.gov

The National Science Foundation Data connected programs have been long-term investments, focused on catalyzing new thinking, paradigms, and practices in developing and using a cyberinfrastructure that meshes data, software and services to understand natural, human, and engineered systems. This presentation will provide an overview of the NSF programs, focused on advanced research cyberinfrastructure. Building a vibrant cyberinfrastructure requires addressing multiple aspects including physical capabilities (high performance computing, networking), software, services, data,

and workforce development. Science and engineering challenges and use cases drive the cyberinfrastructure development, and successful cyberinfrastructure systems strike a balance reflective of both the underlying technology and disciplinary research needs.

Science Use-Cases / **8**

Climate Data: Generating, Accessing and Using Large and Diverse Datasets

Author: Allison Steiner[1]

[1] *University of Michigan*

Corresponding Author: alsteine@umich.edu

Climate scientists conducting research on past, present and future climate use three-dimensional general circulation models to probe questions about the atmosphere and the Earth System. This scientific process generates massive amounts of data covering a broad spatio-temporal range. One example of coordinated climate modeling efforts is the Climate Model Intercomparison Project (CMIP) as part of the Intergovernmental Panel on Climate Change (IPCC) synthesis reports. Generating these data, archiving them in publicly accessible platforms requires a large amount of coordination between global research groups. Accessing this data requires a dedicated amount of training, knowledge of how to use the data, and data storage needs. In this talk, I will discuss climate data from three perspectives: (1) as a climate data generator creating new data to be used in the public domain, (2) as a climate data user who needs to access and analyze climate model intercomparison data, and (3) as a data steward in a role as a journal editor, due to the new requirements by peer-reviewed journals for publicly available data for all publications. Additionally, I will share experiences in teaching data access, usability, and management to graduate students at the University of Michigan who are interested in using and analyzing climate data for a variety of applied climate projects.

Lightning Talks / 11

XSEDE: Collaboration through shared computing resources, data, and expertise

Authors: Gregory Teichert[1]; Brock Palen[None]

[1] *University of Michigan*

Corresponding Author: greght@umich.edu

The Extreme Science and Engineering Discovery Environment (XSEDE) provides an environment for computing resources, data, and knowledge to be shared by researchers worldwide. High-performance and high-throughput computing resources, including large memory nodes and GPUs, as well as data storage and transfer services, facilitate collaborations across multiple institutions for creating and analyzing large data. XSEDE also provides resources to host Science Gateways, which provide higher level interfaces for users to access data, software, and computing resources. Examples of XSEDE Gateways include portals for accessing software for analyzing variability in biological shape, a repository of mathematical formulae, and resources to perform quantum chemistry and molecular dynamics computations. Additional cyberinfrastructure expertise is made available through XSEDE's Extended Collaborative Support Services (ECSS), which supports meaningful collaborations with experts in a variety of areas, including performance analysis, I/O optimization, data analytics, etc. Through these and other resources, XSEDE is a prime tool for the collaboration of scientific researchers.

Related Projects / 9

The Northeast Storage Exchange

Author: Saul Youssef[1]

[1] *Boston University*

Corresponding Author: youssef@bu.edu

The Northeast Storage Exchange is a shared regional storage facility joint project between Boston University, Harvard University, MIT, Northeastern University, UMASS, the Massachusetts Green High Performance Computing Center and RedHat Inc, funded by the National Science Foundation within the DIBBs (Data Infrastructure Building Blocks) program. The NESE team and collaboration aim to create a long-term, growing, self-sustaining storage facility serving both regional researchers and national and international scale science and engineering projects. In fact, the NESE project[3] marks an important shift within the MGHPCC consortium towards shared facilities, shared data science, and much greater collaboration across research computing organizations. We will discuss the state of the initial deployments, buy-ins and relation to growing shared computing facilities in the Northeast.