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

Thursday, 18 May 2017 - Friday, 19 May 2017

University of Michigan

Book of Abstracts

This is the current set of abstracts for the CNSECCS Symposium on May 18-19, 2017.

Contents

Complementary Technology Solutions / **18**

Highly scalable metadata management with signac

Author: Carl S. Adorf[1]

Co-authors: Paul M. Dodd [1]; Sharon C. Glotzer

[1] *University of Michigan*

Corresponding Author: csadorf@umich.edu

Continually increasing computational resources and improved efficiency of parallelized software for data generation and manipulation in the field of scientific computation have led to the requirement of more systematic approaches for data management. We present a data management framework designed to work on both desktop computers and in high-performance computing environments with special emphasis on low entry barriers for both new and experienced users. The *signac* framework assists in the decentralized storage of data and metadata on the file system by providing all basic components needed for building simple to complex data pipelines largely agnostic of data source and format. These managed data spaces are immediately searchable through a homogeneous interface and in this way more accessible to data owners, but also collaborators. Sharing of data across different endpoints is simplified through the generation of metadata indices that contain information about data provenance and current location. The framework's data model is designed not to require absolute commitment to the presented implementation. This reduces barriers for the integration into existing workflows and increases the accessibility to archived data sets. The presented approach simplifies the production of scientific results and collaboration on shared data sets.

Science Use Cases / **10**

Storage challenges in ocean modeling

Author: Brian Arbic[1]

[1] *University of Michigan*

Corresponding Author: arbic@umich.edu

In this talk I will discuss the storage challenges of ocean modeling, using Navy and NASA ocean models as examples.

Science Use Cases / **9**

Slow wave sleep oscillations coordinate neural ensembles during memory consolidation.

Author: Sara Aton[None]

Co-authors: Nicolette Ognjanovski ; Michal Zochowski ; Daniel Maruyama ; Jiaxing Wu

Corresponding Author: saton@umich.edu

The brain routinely integrates polymodal sensory inputs into a coherent representation of events, which is subsequently stored in memory. A long-standing question in neuroscience is how fleeting experiences can modify neural networks to produce memories that are long-lasting, stable, and robust to interference. The advent of new recording technologies allows investigators to monitor and manipulate activity in hundreds or thousands of neurons simultaneously in vivo, during behavior. While this can be used to establish links between neural network activity and brain functions,

two issues complicate this endeavor. First, neural activity patterns typically are quantified over milliseconds-to-minutes timescales, while behaviors evolve over longer timescales (seconds, days, or even years). Second, it is not obvious what features of network dynamics constitute a "signal" associated with a specific brain function, vs. "noise"which is irrelevant to that function.The Aton and Zochowski labs have recently developed metrics to characterize how hippocampal network dynamics change as a function of new learning in mice, during active long-term contextual fear memory formation. We have found that after new information is encoded in hippocampal area CA1 (i.e., following one-trial contextual fear conditioning [CFC]), network dynamics in this area become increasingly stable. This can be demonstrated statistically by calculating changes in mean CA1 network stability (from baseline) across a 24-h period following CFC, and compared this with stability changes in animals which either 1) have undergone CFC, but had subsequent memory formation disrupted through brief post-CFC sleep deprivation (SD), or 2) have undergone a sham behavioral procedure instead of CFC (i.e., where no contextual fear memory is expected). We find that mean stability increases with memory formation, that this increase is disrupted (particularly during slow wave sleep [SWS]) following SD, that these changes are sustained for at least 24 h after learning, and that changes to stability during SWS can predict an individual animal's behavioral contextual fear memory recall 24 h after learning. We also find that the same pattern of network connectivity is consistently repeated for several hours during post-CFC SWS. One feature of SWS which makes it unique from other states is the presence of high-amplitude, low frequency network oscillations in various brain regions. Based on experimental data in which SWS network oscillations in either the hippocampus are transiently disrupted (or mimicked in awake animals), we find that network stability is strongly linked to presence of network oscillations. We propose that stabilization of network dynamics by SWS oscillations could serve as a mechanism to promote long-term memory storage.

Funding Agency Perspectives / **13**

Workflow Science: Moving from tool generation to Discovery

Author: Richard Carlson[1]

[1] *DOE Office of Science*

Corresponding Author: richard.carlson@science.doe.gov

Workflow systems have emerged as the coordination engine behind large distributed data intensive science experiments. They manage the movement of data, the allocation of resources, and display of results for a growing number of science communities. However, existing workflow systems are typically simple, purpose built tools that automate some of the routine tasks a scientist performs. Future workflow systems will need to do more autonomous work, deal with more heterogeneous resources, and provide SMARTer interfaces to both scientists and facilities staff. To achieve this objective Workflow Science needs to move from a tool generation activity to a research and discovery process in its own right. Workflow scientists need to develop the methods, experiments, models, and simulations that can describe and validate the behavior of any workflow system ensuring that it is operating correctly and efficiently.

Science Use Cases / **21**

Neural Interfaces for Controlling Finger Movements

Author: Cynthia Chestek[1]

[1] *University of Michigan*

Corresponding Author: cchestek@umich.edu

Brain machine interfaces or neural prosthetics have the potential to restore movement to people with paralysis or amputation, bridging gaps in the nervous system with an artificial device. Microelectrode arrays can record from hundreds of individual neurons in motor cortex, and machine learning signals can be used to generate useful control signals from this neural activity. Performance can already surpass the current state of the art in assistive technology in terms of controlling the endpoint of computer cursors or prosthetic hands. The natural next step in this progression is to control more complex movements at the level of individual fingers. Our lab has approached this problem in three different ways. For people with upper limb amputation, we acquire signals from individual peripheral nerve branches using small muscle grafts to amplify the signal. After a successful study in animals, human study participants have recently been able to control individual fingers online using acute electrodes within these grafts. For spinal cord injury, where no peripheral signals are available, we implant Utah arrays into finger areas of motor cortex, and have successfully decoded finger flexion and extension with correlations above 0.8. Decoding "spiking band"activity at much lower sampling rates, we recently showed that power consumption of an implantable device could be reduced by 89% compared to existing broadband approaches, and fit within the specification of existing systems for upper limb functional electrical stimulation. Finally, finger control is ultimately limited by the number of independent electrodes that can be placed within cortex or the nerves, and this is in turn limited by the extent of glial scarring surrounding an electrode. Therefore, we developed an electrode array based on 8 um carbon fibers, no bigger than the neurons themselves. We were able to insert arrays with 3x the density of the Utah array by temporarily shortening the fibers for penetration of the top cortical layers. This enabled chronic recording of single units with no apparent contiguous scarring over time. The long-term goal of this work is to make neural interfaces for the restoration of hand movement a clinical reality for everyone who has lost the use of their hands.

Related Projects / **14**

ConFlux: An Ecosystem for Data-enabled Computational Physics

Author: Karthik Duraisamy[1]

[1] *University of Michigan*

Corresponding Author: kdur@umich.edu

The pursuit of accurate predictive models is a central issue and
pacing item in many scientific and engineering disciplines. With the
recent growth in computational power and measurement resolution, there
is an unprecedented opportunity to use data from fine-scale
simulations, as well as critical experiments, to inform, and in some
cases even define predictive models. While the general idea of
data-driven modeling appears intuitive, the process of obtaining
useful predictive models from data is less straightforward.
This talk will discuss a coordinated approach of experimental design, statistical inference and machine learning with the goal of improving the predictive capabilities.

Field inversion is used to obtain spatio-temporally distributed functional terms that directly address discrepancies in the structural form of the model. Once the inference has been performed over a number of problems that are representative
of the underlying physics, machine learning techniques are used to reconstruct the functional corrections in terms of variables that appear in the closure model. When the machine learning-generated model forms are embedded within a standard solver setting, we show that much improved predictions can be achieved, even in geometries and flow conditions that were not used in model training. The usage of
very limited data as an input to construct comprehensive model corrections provides a renewed perspective towards the use of vast, but sparse, amounts of available experimental datasets towards the end of developing predictive models.

The final part of the talk will provide a brief overview of a

hardware/software ecosystem that is being developed at the University of Michigan to enable large-scale data-driven model development for computational physics applications.

Funding Agency Perspectives / 4

Data in the Research Cyberinfrastructure Ecosystem

Corresponding Author: afriedla@nsf.gov

This presentation discusses some of NSF's data programs in the context of advancing the research cyberinfrastructure. This includes interfaces with the physical layer (high performance computing, networking), software, and the research disciplines as well as the potential roles and responsibilities of different stakeholders.

Science Use Cases / 11

Revealing and examining the tempestuous Global Ocean through a multi-petabyte virtual ocean archive.

Author: Chris Hill[1]

[1] *MIT*

Corresponding Author: cnh@mit.edu

This talk will explore how computational science and evolving network and storage capabilities, together with ongoing improvements in remote and in-situ sensing, may be poised, possibly like never before, to have significant impacts on global ocean research. Simultaneous improvements across network, storage, computation and sensing technologies are beginning to create a new lens through which to view, explore and understand some of the key mathematics and observations used to describe and reason about physical, chemical and biological aspects of the Earth's oceans.

Specifically this presentation examines a global one-kilometer horizontal resolution numerical ocean computation that embraces network and storage enabled computational science based approaches. The computation and some of its applications will be described. Some of the key network, storage and computational science technology ingredients that enable the work will be outlined.

The computation examined is work that was recently undertaken using the NASA Pleadies computer. It is one of a new generation of ocean computations that include representations of tidal forcings and realistic synoptic meterology. Including these aspects, at kilometer scale resolution, captures more of the rich dynamics present and observed in the real ocean. This qualitatively increases fidelity of the spatial and temporal variability represented numerically.

Our calculation is initialized from a data constrained estimate of the real-world, large-scale global ocean state. It is driven with boundary conditions taken from high-resolution, data assimilating weather models. The domain is fully global. Interestingly, from a network and storage enabled computational science perspective, we chose to take a uniquely ambitious

approach to storing and distributing the simulation solution. We sampled and archived computation state to a storage subsystem at hourly frequency and at full global resolution for a full year. This created a new and novel resource for ocean research. It is multi-petabyte in size and has global coverage.

The resulting set of more than 10^15 spatially and temporally varying numerical values is supporting a variety of interesting and insightful studies. Many of these would not be easily possible without the underlying network and storage cyberinfrastructure. Advanced cyberinfrastructure underlies archive creation, enables distribution of sizable sub-samples from the archive, and provides tools used in multiple subsequent research studies.

High spatial and temporal storing of the computation more readily reveals an ocean that is teeming with turbulent vorticies and wave motions globally. A series of eye catching visualizations illustrate this. They show what the ocean would look like to eyes that could discriminate components vorticity and density surfaces, instead of visible light!

Examining local regions in frequency wave number space, the stored solution provides notably more complete comparison with theoretical predictions and historical observations than previous generation ocean models. This increased fidelity, combined with the rich sampling archive, is allowing the effort to help guide and support focussed observational field campaigns both at specific locations and globally.

High spatial and frequency capture also allows us to explore new directions in developing statistical relations between readily observable ocean fields and features of interest that are not as directly observable. One example of this, is trying to reduce the stochastic uncertainty due to the ocean internal wave field that impacts acoustic travel time estimates. Underwater acoustics is a potentially powerful tool for measuring the ocean and for creating fully mobile sub-surface networks. It is notoriusly challenging in part because of inherent low bandwidth, but also in part because of the complicated time dependent nature of the ocean as a transmission media. We will illustrate how network and storage enabled approaches can be leveraged in this context. Leveraging these approaches allows us to develop new ways to determine aspects of the internal wave field statistics in a more complete manner. This work draws on the application of statistical methods prevalent in machine learning/big-data communities. Using those methods we can develop various semi-empirical regressions between observable fields and internal wave statistics. Application of these sorts of methods is fundamentally enabled by increasingly robust storage and network cyberinfrastructure technologies.

Another example application looks at the role of high spatio-temporal frequency processes in shaping marine microbial patterns in the ocean. Microbial communities in the ocean form the base of the food chain and play a major, but uncertain, role in Earths carbon, oxygen and nitrogen balance. Marine microbial community structure and ecosystem dynamics remain an area of active research. A highly sampled global fluid solution with spatial and temporal resolution down to scales of kilometers and hours support new ways to explore possible ideas on governing mechanisms for these communties. Recent work in this context will be illustrated.

Finally, we will also sketch briefly the network and storage technologies employed. We will describe approaches for storing data at adequate rates and for disseminating the solution across national networks. The approaches are allowing us to begin to share solutions widely, to local/regional facilities and to cloud services including Dropbox, AWS and Azure. The technical lessons from this exercise show great promise. They provide an illustration of the potential that future ongoing hyperconnected cyberinfrastructure investments could unleash - especially if key technologies are made more routine and

implemented generally in a sufficiently interoperable, capable and cost-effective manner.

Welcome and Overview / 5

Symposium Logistics

Corresponding Author: shawn.mckee@cern.ch

Welcome and Overview / 7

Symposium and CNSECCS Overview

Corresponding Author: shawn.mckee@cern.ch

Related Projects / 20

OSiRIS: Targeting the Multi-science, Multi-institutional Collaborative Data Challenge

Authors: Shawn Mc Kee[1]; Benjeman Jay Meekhof[1]

[1] *University of Michigan (US)*

Corresponding Author: shawn.mckee@cern.ch

We will present an overview of the OSiRIS project (NSF Award #1541335, UM, IU, MSU and WSU) which started in September 2015. OSiRIS's goal is to provide a single scalable, distributed storage infrastructure that allows researchers at any participating campus to read, write, manage and share data directly from their own computing locations even as the data is shared across all participating campuses.

The OSiRIS infrastructure is specifically targeted at addressing the challenges more and more science domains are encountering as they work with large, diverse or distributed data. As the data grows, or becomes more complex or distributed, scientists are challenged to manage, share and analyze that data and become diverted from a focus on their scientific research and instead dealing with data-access and data-management concerns.

We will describe how the OSiRIS project is tackling this challenge using a combination of Ceph, software-defined storage, various open-source management, security and monitoring components and software-defined networking to enable an infrastructure that supports multiple science domains with multi-institutional access to collaboratively extract scientific results from large, distributed or diverse data. The presentation will cover the current status of OSiRIS, describe it's technical details, experiences to-date, and summarize our plans for remainder of the 5-year project.

Science Use Cases / 22

Experiences Sharing Data and Models from a Multi-Institutional Cancer Modeling Consortium

Author: Rafael Meza[1]

[1] *University of Michigan*

Corresponding Author: mcteja@umich.edu

This abstract to be updated shortly

Welcome and Overview / **6**

Welcome Address

Corresponding Author: emichiel@umich.edu

Complementary Technology Solutions / **17**

Bootstrapping Big Data with Spark SQL and Data Frames

Author: Brock Palen[None]

Corresponding Author: brockp@umich.edu

Apache Spark, a popular open source big data tool form the Hadoop ecosystem is seeing rapid adoption across industry and academia, yet it is still generally not well known. For this talk we will demonstrate some large scale samples of how easy it is to benefit form spark SQL and Data Frames for Python and R programmers.

Science Use Cases / **15**

Brain-Life: Engaging neuroscience workforce in big data and reproducible research.

Author: Franco Pestili[1]

[1] *Indiana University*

Corresponding Author: frakkopesto@gmail.com

Neuroscience is engaging at the forefront of science by dissolving disciplinary boundaries and promoting transdisciplinary research. This is a process that, in principle, can facilitate discovery by convergent efforts from theoretical, experimental and cognitive neuroscience, as well as computer science and engineering. To assure the success of this process the current lack of established mechanisms to guarantee reproducibility of scientific results must be overcome. Promoting open software and data sharing has become paramount to address reproducibility. This project addresses challenges to neuroscience reproducibility by providing integrative mechanisms for publishing data, and algorithms while embedding them with compute resources to impact multiple scientific communities.

Complementary Technology Solutions / **16**

Data Ingestion at Scale

Author: Jeffrey Sica[1]

[1] *University of Michigan*

Corresponding Author: jsica@umich.edu

HPC traditionally handles data at rest. The acquisition of streaming data presents a different set of challenges that, at scale, can be difficult to tackle. The approach to building data ingestion infrastructure at ARC-TS involves treating every service as a swappable building block. With this pluggable design using Docker containers you are free to choose which component is best. We will use an example use case to show how data is being generated, ingested, and how each component in the stack can be replaced.

Funding Agency Perspectives / **2**

Big Data –view from National Institute on Aging at the National Institutes of Health

Corresponding Author: silverbergn@mail.nih.gov

Dr. Silverberg will describe some large, NIH and NIA initiatives, such as BD2K, that begin to address big data challenges of current health related research. Additionally, multiple examples of NIH and NIA research projects will illustrate the ever increasing needs for big data solutions. Finally, a few relevant funding opportunity announcements will be shared.

Related Projects / **19**

The NESE Project

Author: Saul Youssef[1]

[1] *Boston University*

Corresponding Author: youssef@bu.edu

Since 2013, Boston University, Harvard University, MIT, Northeastern University and the five campuses of the University of Massachusetts system have jointly housed research computing equipment in a facility in western Massachusetts called the Massachusetts Green High Performance Computing Center (MGHPCC). These same universities have now joined to collaborate on a shared, CEPH-base, MGHPCC-located, regional storage project called NESE (Northeast Storage Exchange). NESE is aimed at providing regional storage for an expanding consortium, providing storage for projects like the U.S. ATLAS Northeast Tier 2 Center, providing storage for Computer Science research projects, and aimed at collaborating with other regional centers in an emerging national cyberinfrastructure. Plans and status of NESE will be discussed.