CIML SI24 Day 2: Prep Day Logistics and Introductions

Mary Thomas

Welcome to the FOURTH CIML Summer Institute!

- Focus is on scalable machine learning.
- GitHub: https://github.com/ciml-org/ciml-summer-institute-2024
- Please be on time so we can stay on schedule.

What is CIML?

- NSF CyberTraining Grant: *Developing a Best Practices Training Program in Cyberinfrastructure-Enabled Machine Learning Research (CIML)*
- Objectives: Scalable Machine Learning
  - To create generalized machine learning training and project materials that run on large-scale NSF funded cyberinfrastructure resources such as XSEDE
  - Targeted towards researchers and educators who are using machine learning (ML) and big data analytics methods for their domain specific applications or instructional material
  - To develop a community of machine learning and data analytics CI Users (CIU) and Contributors (CIC) who actively contribute to the training material repository and incorporate the materials into their projects and courses.
  - Synthesize the training material into a domain independent CIML workflow system that can be used for creating applications that run on the NSF HPC ecosystem.

Logistics

- Friday, June 18th was "Prep Day"
- We focussed on making sure you can connect to Expanse, run jobs, launch notebooks.
- We will use Slack for chatting/communicating
- We will use Zoom for
  - All presentations and group discussions
  - Breakout rooms for hands-on sessions
  - To avoid Zoom fatigue, we'll have lots of breaks
- When speakers have 5 mins left, we will make a clicker sound (demo)
- WebSite: https://na.eventscloud.com/website/22773/home/
- GitHub: https://github.com/ciml-org/ciml-summer-institute-2021

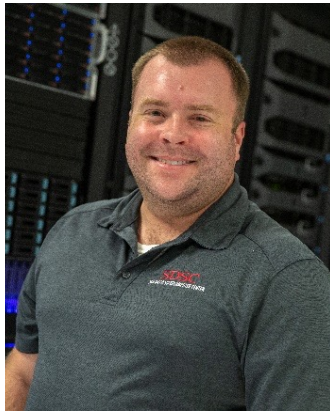
Day2 Agenda: *HPC, Parallel Concepts*

8:00 am -8:30 am	**Light Breakfast & Check-in** Location: SDSC Auditorium
8:30 am - 9:30 am	**2.1 Welcome and Introductions** Mary Thomas, Computational Data Scientist & Director of the CIML Summer Institute
9:30 am - 10:15 am	**2.2 Parallel Computing Concepts** Robert Sinkovits, Director of Education and Training *We will cover supercomputer architectures, the differences between threads and processes, implementations of parallelism (e.g., OpenMP and MPI), strong and weak scaling, limitations on scalability (Amdahl's and Gustafson's Laws) and benchmarking.*
10:15 am - 10:30 am	**Break**
10:30 am - 11: 15 am	**2.3 Getting Started with Batch Job Scheduling** Marty Kandes, Computational and Data Science Research Specialist *Batch job schedulers are used to manage and fairly distribute the shared resources of high-performance computing (HPC) systems. Learning how to interact with them and compose your work into batch jobs is essential to becoming an effective HPC user.*
11:15 am - 12:30 pm	**2.4 Data Management and File Systems** Marty Kandes, Computational and Data Science Research Specialist *Managing data efficiently on a supercomputer is important from both users' and system's perspectives.* *We will cover a few basic data management techniques and I/O best practices in the context of the Expanse system at SDSC.*
12:30 pm - 1:30 pm **Lunch @ Cafe Ventanas**	

1:30 pm - 3:00 pm	**2.5 GPU Computing - Hardware architecture and software infrastructure** Andreas Goetz, Research Scientist & Principal Investigator *Brief overview of the massively parallel GPU architecture that enables large-scale deep learning applications, access and use of GPUs on SDSC Expanse for ML applications*
3:00 pm - 3:15 pm	**Break**
3:15 pm - 4:45 pm	**2.6 Software Containers for Scientific and High-Performance Computing** Marty Kandes, Computational and Data Science Research Specialist *Singularity is an open-source container engine designed to bring operating system-level virtualization to scientific and high-performance computing. With Singularity you can package complex computational workflows --- software applications, libraries, and data --- in a simple, portable, and reproducible way, which can then be run almost anywhere.*
4:45 pm - 5:00 pm	**Q&A, Wrap-up**
5:00 pm - 5:30 pm	**SDSC Data Center Tour**
5:30 pm - 7:30 pm **Evening Reception - UC San Diego, Seventh College, 15th Floor**	

CIML Instructors

Andreas Goetz, Ph.D.
Director of Computational Chemistry Laboratory



Marty Kandes, Ph.D.
Computational and Data Science Research Specialist

Mai Nguyen, Ph.D.
Lead for Data Analytics

Paul Rodriguez, Ph.D.
Research Analyst

Peter Rose, Ph.D.
Director of Structural Bioinformatics Laboratory

Robert Sinkovits, Ph.D.
Director of Education & Training

Mary Thomas, Ph.D.
Computational Data Scientists, HPC Training Lead

Let’s get to know each other

1. Name
2. Institution/Company & Department
3. How do you like to spend your time when not at work?
4. What have you binged watched or read?

Basic Information

- Expanse User Guide:
  - https://www.sdsc.edu/support/user_guides/expanse.html

- You need to have an Expanse account in order to access the system. There are a few ways to do this:
  - Submit a proposal through the XSEDE Allocation Request System
  - PI on an active allocation can add you to their allocation (if you are collaborators working on the same project).
  - Request a trial account, instructions @ https://portal.xsede.org/allocations/startup.

- Online repo and information:
  - https://github.com/sdsc-hpc-training-org/expanse-101
  - https://hpc-training.sdsc.edu/expanse-101/

Resources

- Expanse User Guide
  - https://www.sdsc.edu/support/user_guides/expanse.html
- GitHub Repo for this webinar: clone code examples for this tutorial – clone example code:
  - https://github.com/sdsc-hpc-training-org/expanse-101
- SDSC Training Resources
  - https://www.sdsc.edu/education_and_training/training
  - https://github.com/sdsc-hpc-training/webinars
- XSEDE Training Resources
  - https://www.xsede.org/for-users/training
  - https://cvw.cac.cornell.edu/expanse/

We hope you all have a great workshop!