APS Scientific Computation Seminar Series

Speaker: Roger Harquail French, Kyocera Professor, Materials Science and Engineering
Case Western Reserve University, Cleveland, Ohio

Title: Accelerating Time to Science sans Human Interaction: Materials Data Science, Enabled by Integration of Distributed and High-Performance Computing

Date: October 16, 2023
Time: 1:00 p.m. (Central Time)

Location: Join ZoomGov Meeting
https://argonne.zoomgov.com/j/1601444470?pwd=N1phbHZVdCtmcVR5cGh0c1Zhc0orZz09
Meeting ID: 160 144 4470
Passcode: 937918
One tap mobile
+16692545252,,1601444470# US (San Jose)
+16468287666,,1601444470# US (New York)
Dial by your location
+1 669 254 5252 US (San Jose)
+1 646 828 7666 US (New York)
+1 646 964 1167 US (US Spanish Line)
+1 669 216 1590 US (San Jose)
+1 415 449 4000 US (US Spanish Line)
+1 551 285 1373 US
Meeting ID: 160 144 4470
Find your local number: https://argonne.zoomgov.com/u/af2crdvQy

Hosts: Mathew Cherukara and Nicholas Schwarz

Abstract: Modern materials science research generates petabyte-scale spatiotemporal datasets that span a number of data modalities and formats. Coherent integration of tabular, image, and multimodal graph data is a non-trivial task. We have developed the Common Research Analytics and Data Lifecycle Environment (CRADLE), an analytics infrastructure and framework that supports the scale and diversity of materials science data. CRADLE can handle large-scale, heterogeneous datasets and provides a flexible toolbox for building machine learning pipelines that span from ingestion to model deployment. It is accessible to research scientists with either limited or extensive computational backgrounds and is able to utilize a myriad of low performance to high performance computer systems. CRADLE integrates distributed systems like Hadoop/Hbase/Spark/Ozone with High-Performance Computing (HPC). Materials data scientists can query petabytes of data and train thousands of models in a parallel, distributed environment. We demonstrate four use cases which benchmark its capability to ingest, process, analyze and model spatiotemporal materials data at scale. These tasks span data modalities exemplified in photovoltaics (PV), fertilizer motion through watersheds, and materials characterization and performance investigations. Our applications for tabular data include power forecasting and PV Performance Loss Rate analysis on 29 billion time series power measurements, as well as geospatiotemporal tracking of nitrogen and phosphorus runoffs through watersheds using 29.7 billion elevations. In the case of image data, 4 million XRD diffractograms from high energy synchrotron beamline X-ray diffraction data are used to study in/ex-situ material properties. Multimodal data was utilized to develop spatiotemporal graphs from 3 terabytes of XCT images of Al:Mg stress corrosion cracking to study crack/precipitate interactions. CRADLE accelerates time to science, extends to other domains with similar challenges, and expands the horizon of data science and research.