DS 7347
High-Performance Computing (HPC) and Data Science
Session 2

Robert Kalescky
Adjunct Professor of Data Science
HPC Research Scientist

April 28, 2022

Research and Data Sciences Services
Office of Information Technology
Center for Research Computing
Southern Methodist University

Outline

Session Question

HPC and Data Science

Semester Project

Assignment

Session Question

Session Question

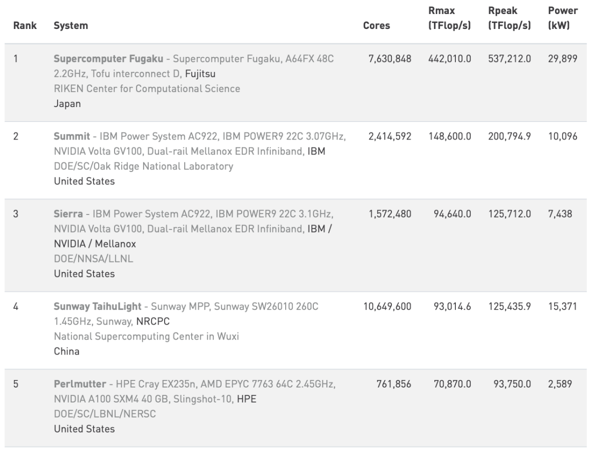
Why does data science at scale need HPC?

HPC and Data Science

Top500

- top500.org
- Twice yearly lists of the top 500 supercomputers in the world
    - Top500, based on HPL performance
    - HPCG, based on HPCG performance
    - Green500, based on HPL performance per watt
- Lists have been kept since 1993

Nov. 2021 Top500 Top Five



Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442,010.0	537,212.0	29,899
2	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148,600.0	200,794.9	10,096
3	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94,640.0	125,712.0	7,438
4	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, NRCPC National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
5	**Perlmutter** - HPE Cray EX235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10, HPE DOE/SC/LBNL/NERSC United States	761,856	70,870.0	93,750.0	2,589

Figure 1: Top five supercomputers.

Fugaku

- COVID-19 Projects
- SALMON (Scalable Ab initio Light-Matter simulator for Optics and Nanoscience)
  - Optimized the simulation computer code to maximize its performance
  - Modeled light-matter interactions in a thin film of amorphous silicon dioxide, composed of more than 10,000 atoms
  - Simulations were carried out using almost 28,000 nodes

Summit

- Air Flow with Classrooms
- Deep Learning to Predict Protein Functions at Genome Scale
  - Datasets of $\sim 45,000$ proteins
  - Use of DeepMind’s AlphaFold 2 for genome sequence to structure converstion

Sierra

- HPC4Mfg project targets more energy-efficient steelmaking
  - Goal is to reduce emissions and defects from inclusions in steel manufacturing
  - Computer vision
  - Machine learning
  - HPC resources
- COVID-19 Risks for Cancer Patients
  - Dataset $\sim 50,000$ patients with cancer
  - Identified a higher risk from COVID-19 due to a specific group of rarer blood cancers and two cancer medications

Sunway TaihuLight

- Millimeter-Scale and Billion-Atom Reactive Force Field Simulation
  - Simulate chemical reactions (ReaxFF)
  - $1,358,954,496$ atoms
  - $4,259,840$ cores
  - Performance of 0.015 ns/day
- Extreme-Scale Earthquake Simulations
  - Simulate Tangshan and Wenchuan earthquakes
  - Leading spatial resolution
  - Memory bandwidth constraints yielded 8% to 16% efficiency

Perlmutter

- November 2020 - July 2021: Cabinets containing GPU compute nodes and service nodes for the Phase 1 system arrived on-site and are being configured and tested
- Summer 2021: When the Phase 1 system installation completed, NERSC started to add users in several phases, starting with NESAP teams
- June and November 2021: The Phase 1 system was ranked at No. 5 in the Top500 lists in June and November, 2021
- June 2, 2021 and January 5-7, 2022: User trainings were held to teach NERSC users how to build and run jobs on Perlmutter
- January 19, 2022: The system is available to all users who want to use GPUs with the start of the allocation year 2022

Semester Project

Semester Project Goal

Produce single-submit, end-to-end, performant pipeline for a complex and computationally intensive data analysis workflow.

Semester Project Details

- The analysis and dataset, possibly generative, needs to be sufficiently computationally intensive such that a resonable performance analysis can be conducted.
- The specific dataset, analysis, and performance analysis will be agreed to at various stages during the semester.
- The pipeline should be single-submit, meaning that a single job is submitted to the queue system and then entire pipeline is run with each stage run on appropriate hardware with appropriately optimized software stacks.

Semester Project Details

- The deliverable will be a ready to present slide deck in your GitHub repo, *i.e.* a job will be submited on an SMU HPC cluster and then, sometime later with zero human interaction, a PDF presentation will appear in your GitHub repo.
- The presentation should discuss both the dataset analysis and performance analysis.
- Specific compute resources will be reserved for final testing and the production run.

Finding Datasets

- ERNESTO.net
- KDnuggets
- Data Is Plural
- Open Data on AWS

Group Discussion

- Discuss ideas
- Find some interesting datasets

Assignment

Assignment 2

Readings

- Eijkhout sections [1.0–1.4]

Assignment

- Find three HPC and Data Science example articles or papers using three, non-top five of the Top500, supercomputers (can be decomissioned machines)
- Briefly discuss the research workflow, specific hardware used, and cite the sources
- Commit `assignments/assignment_02.md` to your class repo
- Due 12:00 AM Central, Tuesday, May 3, 2022