Spark @ NERSC

Prabhat[1], Michael Mahoney[2], Jey Kottalam[2], Jiyan Yang[3], Mike Ringenburg[4], Venkat Krishnamurthy[4]

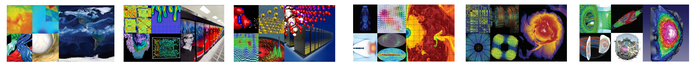
[1]NERSC, [2]AMPLab, [3]Stanford University, [4]Cray, Inc

NERSC

The National Energy Research Scientific Computing Center (NERSC) at Lawrence Berkeley National Laboratory is the Primary High Performance Computing & Data Facility for DOE Office of Science Research.

- Over 5000 users from across the world
- 1500 scientific publications annually
- 3 recent Nobel laureates
- Wide range of scientific applications

NERSC is facing a data deluge! Data is becoming increasingly important as petascale and exascale class simulations, as well as observational and experimental modalities of science produce massive datasets.



Bio Energy, Environment | Computing | Materials, Chemistry, Geophysics | Particle Physics, Astrophysics | Nuclear Physics | Fusion Energy, Plasma Physics

Top 10 Analytics Problems

1. Creating a catalog of all objects in the visible universe
- Joint inference across all telescope data

2. Determining the fundamental constants of Cosmology
- Clustering and n-point correlation on Trillion particle datasets

3. Characterizing extreme weather in a changing climate
- Pattern detection on PB-sized climate datasets

4. Knowledge extraction from scientific literature
- Natural Language Processing, Q&A, Hypothesis generation

5. Understanding speech production
- Machine Learning for decoding neural activity

6. Cellular imaging
- Raw images stacks to semantic databases

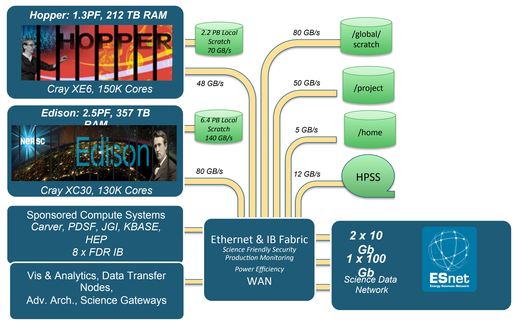
7. Mass-spectroscopy imaging
- Matrix factorizaton, clustering for O(M) dimensional points

8. Understanding the genetic code
- Sequencing, Assembly

9. Material Genome
- Predicting material properties from sparse datasets

10. Fundamental constituents of matter
- Pattern/Anomaly detection for LHC, Daya Bay, LUX

Computational Systems



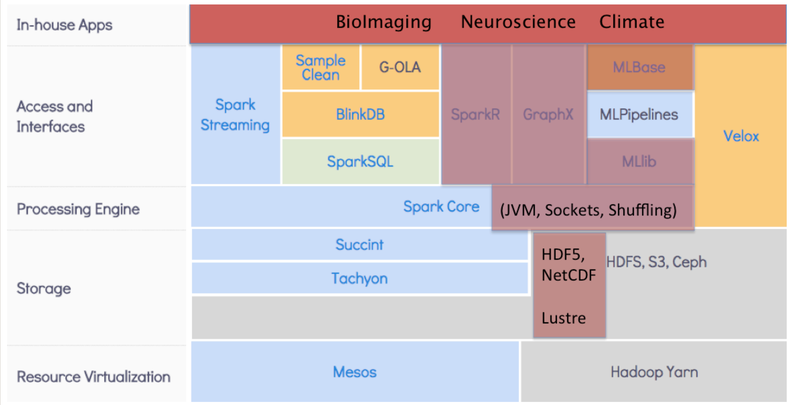
- Cori will transition HPC and data-centric workloads to energy efficient architectures
- 9,300 'Knights Landing' compute nodes with Aries Interconnect
- 28 PB Lustre File system, >700 GB/sec peak performance
- Data partition features (1900 Haswell Nodes):
  - NVRAM "Burst Buffer" for I/O acceleration
  - Large memory nodes
  - Massive serial jobs
  - Complex workflows
  - Communication with external services
  - Streaming Data from observational facilities
  - Easy to customize environment (Containers)

Big Data Software Portfolio

Capabilities	Technology Areas	Tools, Libraries
Data Transfer + Access	Globus, Grid Stack, Authentication	Globus Online, Grid FTP
	Portals, Gateways, RESTful APIs	Drupal/Django, NEWT
Data Processing	Workflows	Swift, Fireworks
Data Management	Formats, Models Databases	HDF5, NetCDF, ROOT MongoDB, SciDB, PostgreSQL, MySQL
	Storage, Archiving	Lustre/GPFS, HPSS, SRM
Data Analytics	Statistics, Machine Learning	python, R, ROOT BDAS/Spark
	Imaging	OMERO, Fiji, MATLAB
Data Visualization	SciVis InfoVis	VisIt, Paraview Tableau

Spark provides a productive interface for NERSC users interested in data analytics. We have installed and created a module for Spark at NERSC. APIs such as MLBase, SparkR and GraphX are of particular interest.

BDAS on Cray XC systems

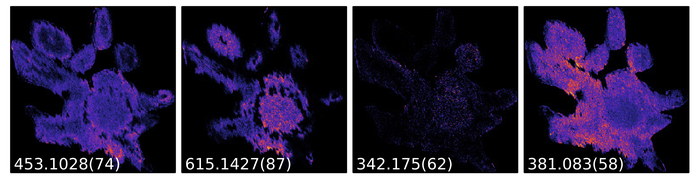


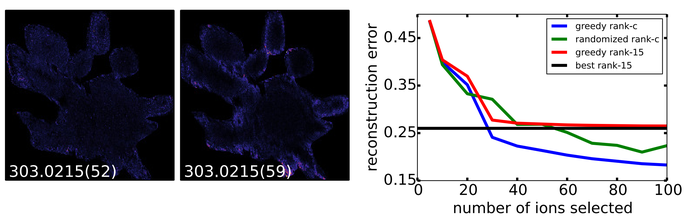
A new collaboration between AMPLab, Cray and NERSC is taking a closer look at performance of the BDAS stack on Cray XC30/XC40 class HPC systems.

We are considering a range of data-centric workloads from BioImaging, Neuroscience and Climate. Matrix factorization (PCA and CX) are important analytics methods in these domains, and we are extending the linear algebra capabilities in BDAS stack to operate on TB-sized datasets.

Application to BioImaging

We are implementing CX/CUR methods in Spark, and stress testing them on large scale mass-spectroscopy imaging datasets. Figures below demonstrate the successful application of CX decomposition to a drug producing shrub (Podophyllum hexandrum) provided by Norman Lewis from Washington State University. This is the first time that matrix decomposition has been attempted for ion mobility datasets.





U.S. DEPARTMENT OF ENERGY | Office of Science