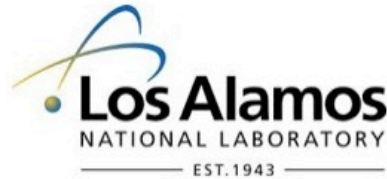




Lawrence Berkeley
National Laboratory



HPCOR 2014

User Training for Data Intensive Science
Co-Chairs: Fernanda Foertter, Tim Fahey

Contributors

- Blaise Barney, LLNL
- Karen Haskell, SNL
- Bob Balance, SNL
- Fernanda Foertter, ORNL
- Tim Fahey, LLNL
- Richard Gerber, NERSC
- Dee Magnoni
- Glenn Lockwood
- Ashley Barker; K. Fagnan, NERSC
- David Karelitz; Richard Coffee ANL

What are your major strategies and initiatives over the next 5 & 10 years? How do they affect staffing levels?

- Movement between data levels: memory, L1, L2, disk, tape.....
- Sharing/Leveraging training efforts at other laboratories; don't reinvent the wheel
- Training on data management plans and curation of data
- DMP (sponsor requirements) vs. (user expectations) vs. (site capabilities)
- Transforming data to structured format
- Training on workflows
- I/O; Is my I/O bad? Why is it so slow? Profile first? Tools for profiling I/O are needed. Tool's for monitoring I/O (compile time from each file system)
- Training by exposing system data (for lscratch, pbatch)

What are your current efforts and/or site configuration in this area?

- SDSC/NFS tells users they are required to have a data management plan but don't provide a storage site for them. The data must be hosted off-site. NSF is now auditing Authors on their data management commitment.
- Dmptools: Sherpa Juliet, dmp.cdlib.org gives templates for DMP's.
- Preserving the discovery process of evaluated tools?
- Can we build an trusted expertise database
- Can we build a Wiki

What are your mandates and constraints?

- Common site for all DOE labs/users to put user documentation and feedback. Information is shared between the labs, not public. DOE bubble.
- Protecting data vs. sharing data; NNSA labs must break out of habit of holding data close.
- JGI training; HPC 1-week intro; inviting external expertise (python)
- Software carpentry; software development best practices
- Add a Catalyst, Consultant

How to do you forecast future needs and requirements?

- Can we develop and share training?
- Create and aggregating modules?
- How much effort do I devote to training our users?
- New getting started workshops at ANL; 6 users maximum, 90 minutes dedicated; not drop in-drop out of a web site.
- Some users want to do homework.
- Leverage each sites expertice (BG/Q, GPU's)
- Hackathon for training? Need to bring in expertise to monitor the hackathon.

What opportunities exist for productive collaborations among DOE HPC centers?

- Leverage expertise at each site.
- NERSC has procmon for monitoring file system health and performance
- Q. How do we determine the outcome of our training or the effectiveness (say 6 months later)
- Is a drivers license required? Should one be?
- Tell us something we did right; tell us something we did wrong; training metrics: collect review and follow up.
- User meetings: invite a speaker to get attendees.

What are the biggest challenges and gaps between what you can do today and what will be required in 5 - 10 years?

- Teaching scientists and researchers how to tag their data with metadata
- Digital Object Identifier
- Creating a data working group to discuss the data
- Research Library
- Coursera on

Describe some practices that you think are effective as well as lessons learned that would be helpful to other centers?

- Blah
- Blah
- Blah

Top Findings

Opportunities

- Learn of and take advantage of each others expertise.
- Identify what steps we could we take to centralize our training. Who would own and host TBD; what are security concerns we need to address.
- What training is available for all users at all labs. We just don't know.
- Vis, workflow, metadata and I/O focused topic areas are areas for training improvement.
- Web Ex/ Blue Jeans virtual training
- Data Management Plan requirements
- SW Carpentry for SW dev Best Practices

Best Practices

- Your strengths relative to competitors. Wh's best at gpu's. MPI, BB,
- Training users via exposing system data on file system performance
- Training feedback loop between developers, users and policy
- Monthly virtual meetings in anticipation of delivering virtual training (ORNL)
- Collaboration between sites developing common workshop materials for ATS-1 and ATS-2 procurements
- Understanding the roles and responsibilities of all levels of the projects (PI's vs PM's vs Post Docs..)

Challenges

- Adapting resources to the changing requirements of our users.
- Develop training for ATS systems and other new technologies
- Data Curation/Data Management Plan
- Need to be kept informed of each others efforts in user training
- Developing focusing on workflow management tools (hadoop, firefly?)
- Burst buffer, nvram, L1 and L2 cache utilization undefined (developer vs user?)
- Access to test beds of newest technologies to train the trainers
- Utilize StackOverflow?