Management, Analysis, and Visualization of Experimental and Observational Data
The Convergence of Data and Computing

Advanced Scientific Computing Advisory Committee Meeting
9/20 – 9/21/2016
Washington, D.C.

E. Wes Bethel, Ph.D.
Lawrence Berkeley National Laboratory

EOS: Multiple Exabytes of Data/Year

- Detectors and other sensors increasing in resolution and speed faster than processors and memory.
- Science User Facilities (SUFs) are estimating O(10s) PB/yr
- Across SC SUFs in the coming years, the aggregate forecast is multiple EB/yr.
- Are we ready?

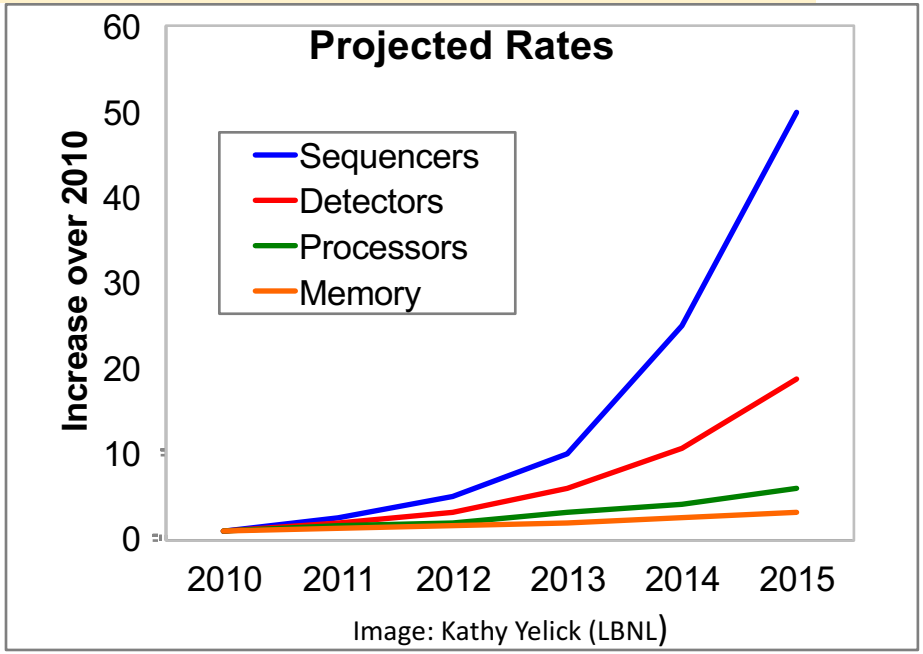
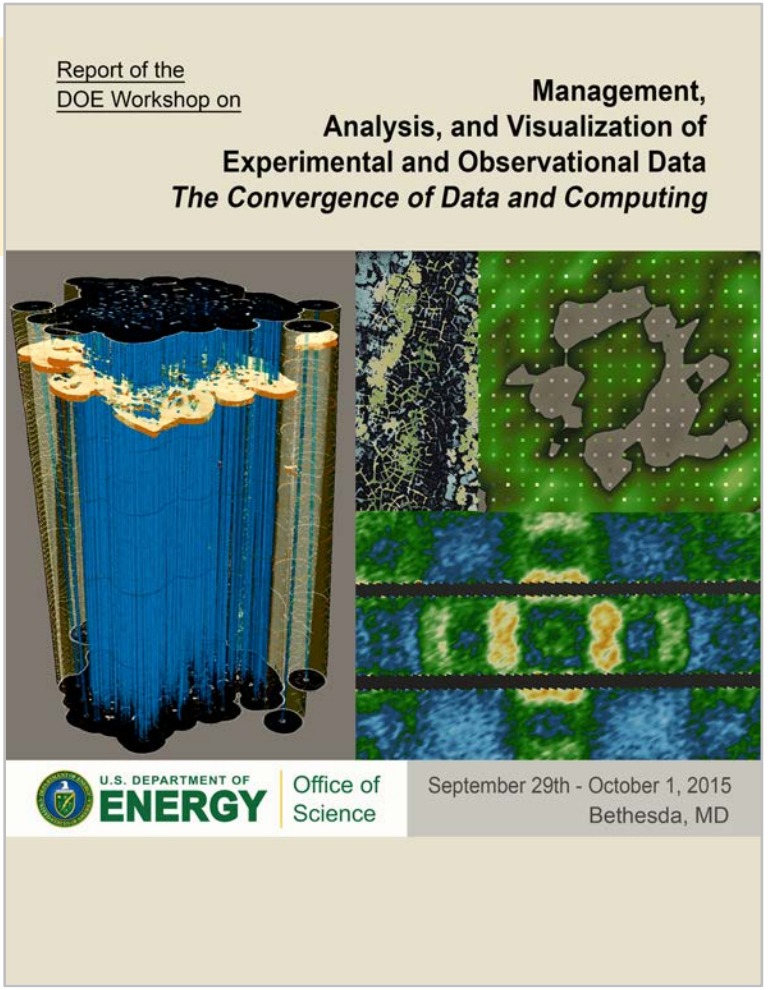


Image: Kathy Yelick (LBNL)

U.S. DEPARTMENT OF ENERGY | Office of Science

EOD Workshop Objectives

- Better understand data-centric issues facing Office of Science Science User Facilities
- Identify for meeting those needs and science objectives
- Foster dialogue between EOS projects and ASCR
- September 29 – October 1, 2015



U.S. DEPARTMENT OF ENERGY | Office of Science

Who Was At the Workshop?

Science User Facility	Lab	Office
Environmental Molecular Sciences Lab	PNNL	
Climate Modeling (Computing)		
Atmospheric Radiation Measurement Climate Research Facility		
Advanced Light Source		
Linac Coherent Light Source		
Spallation Neutron Source High Flux Isotope Reactor		
Scanning Tunneling Electron Microscopy	ORNL	BES
		BES

Science User Facility	Lab	Office
Deep Underground Neutrino Experiment	FNAL	HEP

Math/CS/Facilities - ASCR
Computing/Network Facilities: NERSC, OLCF, ALCF, ESnet
CS/math: data management, data analysis (ML, graph analytics, statistical analysis, etc.)/visualization, data mining, operating systems/runtime, workflow, optimization, UI/human factors, applied mathematics

Guests
NSF Computational Facility
UK Science Grid

Executive Summary

- Gaining scientific knowledge from experimental data is increasingly difficult
- Convergence of data and computing: data- and computing-centric needs increasingly intertwined, symbiotic
- Acute, urgent data-centric needs in SUFs and science programs

Finding: All EOS Projects Struggle with a Flood of Data

- How much data? ***Multiple exabytes/year***
- Science driver: better instruments
- Challenges:
  - More volume, complexity, variety (the V's)
  - Impediments to understanding data
  - Inadequate infrastructure, software
- Opportunity cost: potential loss of science

A key limitation today is our [in]ability to analyze and visualize the acquired data due to its volume, velocity, and variety.

B. Toby (ANL)
Advanced Photon Source
X-ray imaging/microscopy

Finding: EOS Projects Have Difficulty Using Large-Scale, High Performance Computing Facilities

- Insufficient resources: compute, storage, software, staffing, …
- EOS workloads/needs are different than traditional HPC workloads:
  - Computational requirements
  - Data requirements
- Challenge: many HPC facilities' operational policies optimized for high-concurrency, batch workloads

Procedures for moving data from place to place, including tools for automating resilient workflow for orchestrating distributed data-related operations are a bottleneck.

P. Rasch (PNNL), Climate modeling and analysis

Finding: EOS Projects Have Unmet Time-Critical Data Needs

- EOS projects require low-latency, high-throughput response
- Drivers:
  - (Convergence topic) Experiment optimization/tuning: near real-time analysis of experiment results (using HPC platforms) to adjust experiment parameters to obtain better data.
  - Increasing throughput to keep pace with data acquisition rates.
- Challenges:
  - Local vs. remote computing resources
  - Solutions need to accommodate both local and remote resources
  - Algorithms and software infrastructure designed for workloads of the last decade (or last century) likely inadequate to accommodate these data loads and response times.

Predicting optimal [experiment] parameters could optimize data collection schemes and ultimately provide better quality data.

B. Toby (ANL)
Advanced Photon Source
X-ray imaging/microscopy

Finding: The Risk of Unusable Data

- Without adequate metadata, scientific data has limited usefulness
- Data-centric tasks left to the individual user
- Drivers
  - Compliance with Executive Order for data dissemination
  - Opportunity for new science unforeseen by original experimentalists when data is shared, reused
- Challenge: metadata is an afterthought

One very real problem is that data is almost never usable by anyone other than the person who produced it. This problem must be solved if making data publicly available is to have any useful purpose.

H. Steven Wiley (PNNL)
Environmental Molecular Sciences Laboratory

Finding: Collaboration and Sharing are Central to EOS Projects

- Sharing and collaboration are central to modern EOS
- Driver: Share data, methods and tools with the broader scientific community
- Challenge: *Ad hoc* approaches impede sharing
- Opportunities
  - Reduce costs, better science

Current technologies are inadequate for sharing data between group members. The community needs a more fluid means for sharing data and working together.

H. Steven Wiley (PNNL)
Environmental Molecular Sciences Laboratory

U.S. DEPARTMENT OF ENERGY | Office of Science

Finding: EOS Data Lifecycle Needs Not Being Met

- EOS data lifecycle needs are complex
  - DAQ through data publication and data products
- Drivers:
  - EOS mission is to collect data and share it
  - Reference datasets
- Challenges:
  - No program-wide approach for meeting needs
- Opportunities
  - Potential to increase science discoveries per experiment through data sharing, reuse (reference datasets)

..our only archival process right now is that provided by the published journal.

Providing more access to data, in a manner that can be used by more scientists, will improve efficiency, increase scientific impact, and result in more discoveries per experiment.

G. Granroth and T. Proffen (ORNL)
Spallation Neutron Source

Finding: Software Has a Central Role in EOS Projects

- Software central in all aspects of working with EOD
- Drivers:
  - EOS vulnerable to inefficiencies, increased costs that can result from software-related issues
- Challenges:
  - How to create the scientific software needed to run the science facility?
  - Insufficient program-wide visibility, coordination
- Opportunities
  - Potential for broad, positive impact on EOS

The most serious impediment the APS encounters is a lack of a DOE-wide view of software needs across the BES mission. Since each lab has its own portfolio of responsibilities, it devotes resources to those goals

B. Toby (ANL)
Advanced Photon Source
X-ray imaging/microscopy

Finding: Workforce Development and Retention Concerns

- Our personnel: our single most precious resource
- Challenges:
  - Specialized and multi-disciplinary knowledge is rare
  - Data scientists in high demand
  - Insufficient or inadequate career paths
- Opportunities:
  - EOS is a problem-rich environment

Many data-centric problems require close interaction between the domain scientist and data scientist. Generally, such dual background is the exception and some amount of professional training is required to fill this gap.

S. Kalinin (ORNL)
Scanning probe and scanning transmission electron microscopy

Recommendation: Address Challenges Posed by Growing Data Size, Rate, Complexity

- Data infrastructure: modernize data infrastructure to accommodate present and future EOS data size, rates
- Math and CS research: multidisciplinary, and targeting key EOS data challenges
- Software: cultivate multidisciplinary teams focused on EOS software development and deployment

A key limitation today is our [in]ability to analyze and visualize the acquired data due to its volume, velocity, and variety.

B. Toby (ANL)
Advanced Photon Source
X-ray imaging/microscopy

Recommendation: Evolve HPC Facilities to Include Focus on EOS Needs

- Identify and prioritize EOS needs
- Evaluate strategies for accommodating EOS project needs (convergence, data lifecycle)
- Refine operational policies for EOS workloads
- Reconsider facility metrics, hardware & software implications
- Consider approaches for providing resources, and use policies, attuned for EOS data needs

Need for community-(or facility-)centric data repository for data archival, sharing; with substantial bandwidth to the stored data, and easy interface for interacting with the data analytics… needs to be massively parallel, a combination of visualization and various analysis tools.

Alex Szalay
Johns Hopkins University

Recommendation: Develop Solutions for EOS Data-Centric Workloads

- Assess and understand EOS needs, use patterns, performance limits
- Assess applicability of related factors
- Develop requirements, strategies, policies aimed at EOS workloads
- Cultivate and adopt new methods for specialized EOS use cases like time-critical processing

> . . . it is getting to point where users cannot just download their data—their hard drive isn't big enough, and if it was they wouldn't have the computing power needed to do anything with it.
>
> *D. Parkinson (LBNL) Advanced Light Source*

Recommendation: Improve EOS Productivity with Resilient, Automated Data Pipelines

- Develop understanding of common patterns in EOS
- Develop solutions for optimizing across both quantitative and qualitative metrics
  - E.g., performance, ease-of-use
- Promote reusability of workflows and methods across EOS projects and HPC facilities

Many instruments produce very large and complex data streams that remain unmined … and are quickly becoming the largest fraction of our data by volume … because they must be manually interpreted by experts for data quality and meaning.

L. Riihimaki and C. Sivaraman (PNNL) Atmospheric Radiation Measurement Climate Research Facility

ENERGY | Science

Recommendation: Address Metadata Needs of the EOS Community

- How is data to be used by EOS?
- Focus R&D methods to enable sharing, understanding, search, curation and related EOS needs
- Strive towards automated metadata capture
  - Make it “part of the culture, part of the practice”
- R&D&D of tool sets for capturing, storing, managing metadata

Truly reusable data requires a significant amount of associated metadata…the overall cost and complexity of metadata recording and consolidation is currently prohibitive, which is the primary reason it is rarely collected.

H. Steven Wiley (PNNL) Environmental Molecular Sciences Laboratory

Recommendation: Expand Capabilities for Collaboration and Sharing of EOS Tools & Data

- How do EOS projects need/want to share data?
- Develop approaches for sharing data, finding data, tools, and related resources
- Deploy tools that are data-/metadata-driven, and easy to use

Providing more access to the data, in a manner that can be used by more scientists, will improve efficiency, increase the impact of the science, and result in more papers per experiment.

G. Granroth & T. Proffen (ORNL)
Spallation Neutron Source and High Flux Isotope Reactor

Recommendation: Identify and Fill Gaps in Data Lifecycle, Reproducibility and Curation

- Develop understanding of needs/uses, assess what's available
- Develop a roadmap for R&D that is broadly applicable across SC EOS projects
- Develop strategies for provisioning infrastructure for data needs
- Assess possibility of provisioning DOE-SC-wide facilities for data storage, archival, sharing

A welcome addition in the data universe would be a centralized DOE facility that provides … a mechanism for data archival and retrieval.

B. Toby (ANL)
Advanced Photon Source
X-ray imaging/microscopy

Recommendation: Expand Efforts Focusing on EOS Software Ecosystem

- Develop and understanding of broad and diverse EOS data-centric software needs
- Cultivate software R&D projects for EOS that follow best practices and is broadly applicable, usable
- Identify opportunities for broader coordination of EOS data software design, development, deployment
- Identify and establish practices for software archival, curation, dissemination, and long-term support.

The most serious impediment the APS encounters is a lack of a DOE-wide view of software needs across the BES mission. Since each lab has its own portfolio of responsibilities, it devotes resources to those goals

B. Toby (ANL)
Advanced Photon Source
X-ray imaging/microscopy

U.S. DEPARTMENT OF ENERGY | Office of Science

Recommendation: Develop and Nurture a Data Science Workforce

- Prioritize the role of data science activities in multidisciplinary teams
- Recognize, reward the roles and skills of this group
- Provide career paths

> … Lack of trained manpower and career paths for computationally-oriented scientists
>
> *S. Habib (ANL)*
> *Cosmic Frontier Use Cases*

Convergence: EOS and Simulation Science Increasingly Intertwined and Inseparable

- Better computing results in better data, better data results in better computing
- Diversity of EOS use cases, ”computing” refers to simulation science, and a lot more
- Can’t do data-intensive science without computing

Additional Topics - Backup Slides

- Methodology for collecting SUF data needs
- What happened at the workshop
- EOS Science Use Cases

Summary and Final Thoughts

- All EOS projects face significant, complex data challenges
  - They are tackling these issues “on their own”
  - There is there is no program-wide, coordinated effort to address them
- Urgent and acute needs: multiple exabytes/year are coming soon and EOS projects are not ready
- The workshop report has a lot more depth


The End