THE 2018 DOE ASCR WORKSHOP ON STORAGE SYSTEMS AND I/O

Held Sept. 20, 2018 in Gaithersburg, MD

Rob Ross
Mathematics and Computer Science Division
Argonne National Laboratory
rross@mcs.anl.gov

ASCR's Many Contributions in Storage and I/O

- **Parallel file systems and I/O middleware**
  e.g., DeltaFS, PVFS, ROMIO, PLFS, SCR, VeloC
- **Scientific data libraries and frameworks**
  e.g., ADIOS, HDF5, FlexPath, Parallel netCDF
- **File system alternatives**
  e.g., DataSpaces, Mochi, Proactive Data Containers
- **Indexing and search**
  e.g., FastBit, FastQuery, ALACRITY, Giga+, IndexFS
- **Understanding I/O systems**
  e.g., CODES, Darshan, Pablo, TOKIO

The R&D 100 Awards recognize the 100 most innovative technologies each year. ASCR has supported numerous storage and I/O technologies that have received this prestigious award, including ADIOS, Darshan, FastBit, HDF5, and ROMIO.

See https://www.rd100conference.com/

SSIO Workshop Organizers

Co-Organizers: **Rob Ross and Lee Ward**
ASCR Point of Contact: **Lucille Nowell (now Laura Biven)**

Organizing Committee:
- **Gary Grider**
- **Scott Klasky**
- **Glenn Lockwood**
- **Kathryn Mohror**
- **Brad Settlemeyer**

ASCR Research Division Admin:
Angie Thevenot

Pre-Workshop Report Contributors:
- **Phil Carns**
- **Quincey Koziol**
- **Matthew Wolf**

ORISE Workshop Coordinator:
Deneise Terry

SSIO Workshop Charge

As HPC architecture becomes more complex, **the lines between what operating and runtime systems experts call *memory* and the emerging off-system storage hierarchy that includes solid state devices blur**. These changes result in increased complexity for application developers and increased difficulty in managing the entire process for input and output. A combination of **rapid change in memory and storage technology** and meeting the related requirements for the **range of application classes using high performance computing** (HPC) must drive the prioritization of essential new research activities in the SSIO area. The goal of this day-and-a-half workshop is to identify technical requirements and basic and advanced research directions that will advance the field over the next **5-7 years**.

Organizing, Storing, and Accessing Data for Scientific Discovery (5-7 years time to impact)

Drivers

- Scientific
  - Increasing need to support big data and learning applications
  - Rapid growth of scientific dataset sizes
- Technological
  - New, solid-state storage and tight integration in platforms
  - New accelerators, sensors, and networks

Key Questions

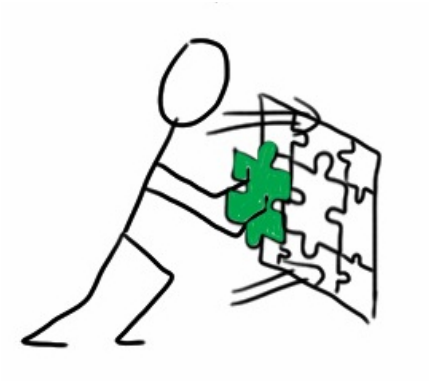
- How do we maintain scientists' productivity while leveraging complex and multi-layer storage systems?
- How can AI/ML assist in managing complex storage environments?
- What new software will be needed to adapt to streaming data sources?
- How do we enable storage to cooperate with workflow and scheduling systems?
- How do we motivate and maintain user trust?

Research Opportunities

- Enabling science understandability and reproducibility through rich data formats, metadata, and provenance
- Accelerating scientific discovery through support of in situ and streaming data analysis
- Enhancing SSIO usability, performance, and resilience through monitoring, prediction, and automation
- Improving efficiency and integrity of data movement and storage through architecture of systems and services


DRIVERS AND QUESTIONS

Identifying Opportunities for Research



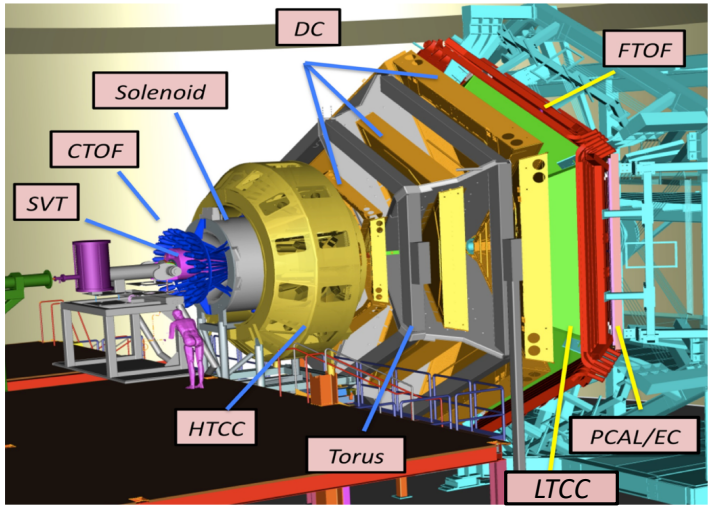
Application pull:
- What do our scientists and facilities need?
- Research solutions to fill those gaps

Technology push:
- What new technologies might be beneficial?
- Research how to employ those technologies productively

Image from M. Geurden, “Market Opportunity Identification: Push or Pull?,”July 2012, https://newentrepreneurship.nl/2012/07/02/market-opportunity-identification-push-or-pull/

Application Pull

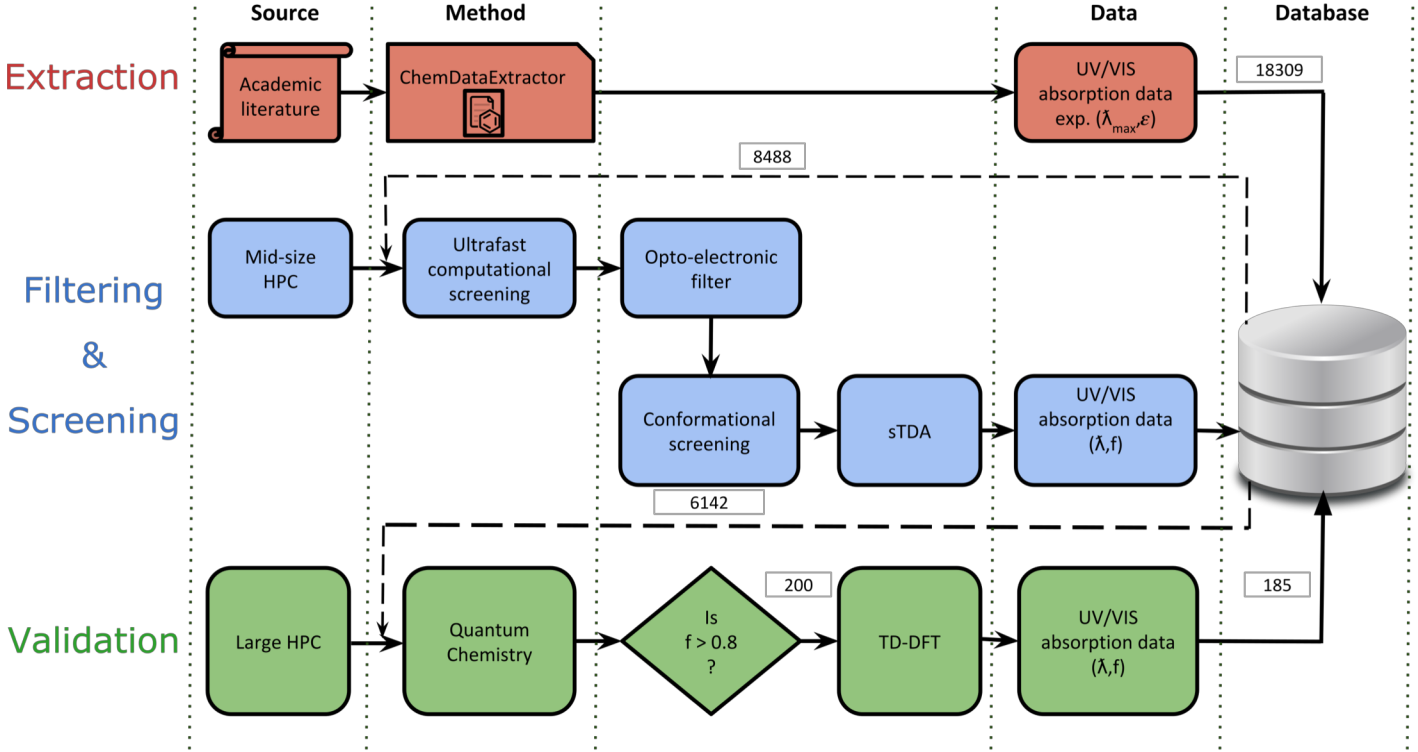
- Experimental/observational data science
  - **Streaming data model**
  - Many small records
- Learning applications
  - **Unstructured data**
  - Random access to large datasets
- In situ data analysis
  - **Fine grained sharing between tasks**
  - High penalty for data transformations
- Reproducibility
  - **Greater reliance on provenance information**



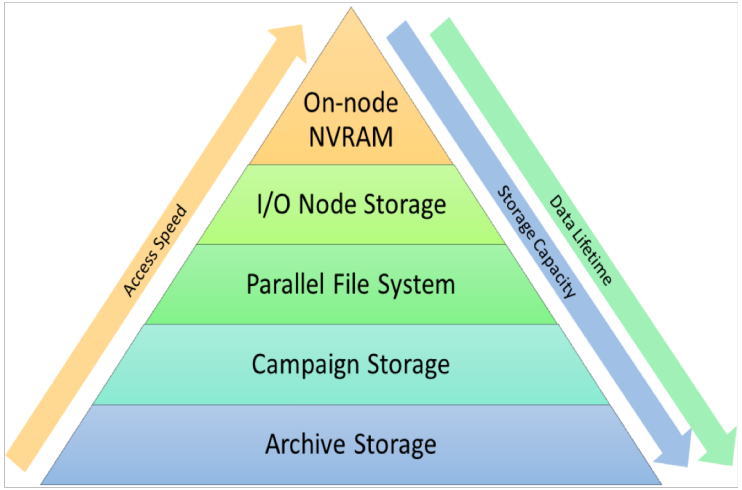
The CLAS12 detector at Jefferson Lab has hundreds of thousands of data sources, generating tens of PBytes of raw data each year. A complex series of steps are performed in near real time to accomplish the first steps of analysis.

Support for streaming data analysis could dramatically accelerate time to scientific discovery. *Image credit G. Heyes (JLab).*

Storage and I/O in Materials Design



G. Sivaraman, "UV/Vis absorption spectra database auto-generated for optical applications via the Argonne data science program," APS March Meeting, March 4, 2019.



New memory and storage technologies provide opportunities to retain more data than ever before, to directly and efficiently access individual records regardless of location in the system, and to lower costs by employing the most economically viable technologies for specific tasks.

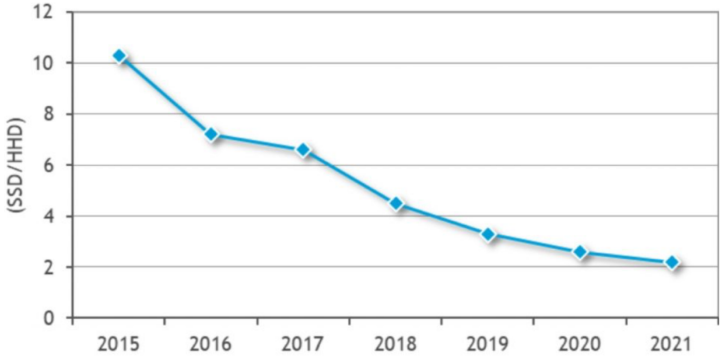
These technologies also change the ways in which storage, workflow, scheduler, and operating systems must work together.

Technology Push

- New memory and storage technologies
  - **Blurring lines between storage and memory**
  - New access methods
- High degree of concurrency from embedded storage devices
  - **High cost for global coordination**
  - New scale and environment for faults
- Deeper storage hierarchy than in the past
  - **Positioning and locating data more difficult**
  - Widely varying performance characteristics
- Interconnects with new characteristics
  - **Emerging quality of service features**

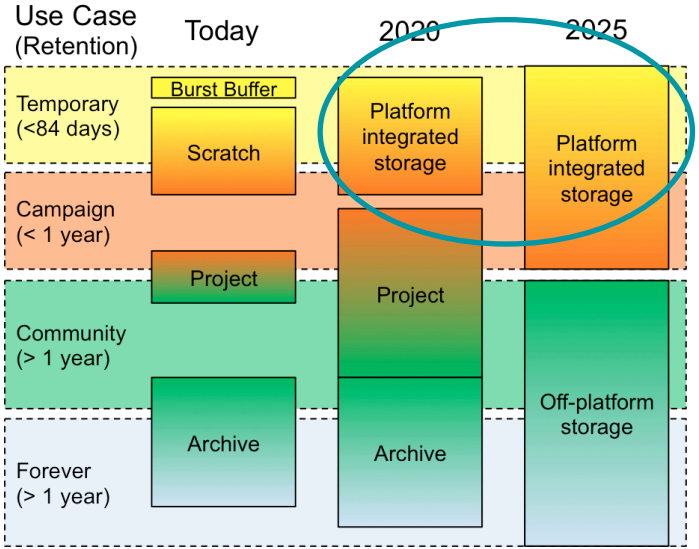
New Uses for Existing Tech

Solid-state disk vs. hard disk drive pricing (per GB ratio)



Source: Hyperion research
https://www.storagenewsletter.com/2018/08/07/flash-storage-trends-and-impacts

Evolution of the NERSC storage hierarchy between today and 2025



Continued decline in cost of SSD capacity relative to HDD has led to plans to employ SSD-backed platform storage, integrated into the platform.

G. Lockwood et al. "Storage 2020: A Vision for the Future of HPC Storage," October 2017, https://escholarship.org/uc/item/744479dp

Questions on the minds of the attendees

Drivers

- Scientific
  - Increasing need to support big data and learning applications
  - Rapid growth of scientific dataset sizes
- Technological
  - New, solid-state storage and tight integration in platforms
  - New accelerators, sensors, and networks

Key Questions

- How do we maintain scientists' **productivity** while leveraging complex and multi-layer storage systems?
- How can **AI/ML** assist in managing complex storage environments?
- What new software will be needed to adapt to **streaming** data sources?
- How do we enable storage to cooperate with **workflow** and scheduling systems?
- How do we motivate and maintain user **trust**?

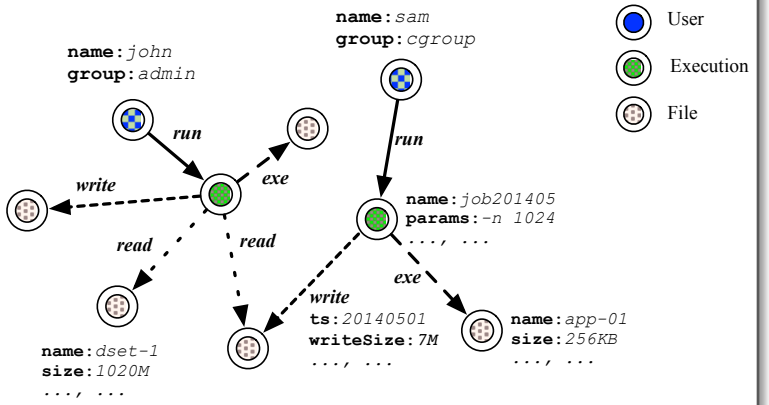
Research Opportunities

- Enabling science understandability and reproducibility through rich data formats, metadata, and provenance
- Accelerating scientific discovery through support of in situ and streaming data analysis
- Enhancing SSIO usability, performance, and resilience through monitoring, prediction, and automation
- Improving efficiency and integrity of data movement and storage through architecture of systems and services


RESEARCH OPPORTUNITIES

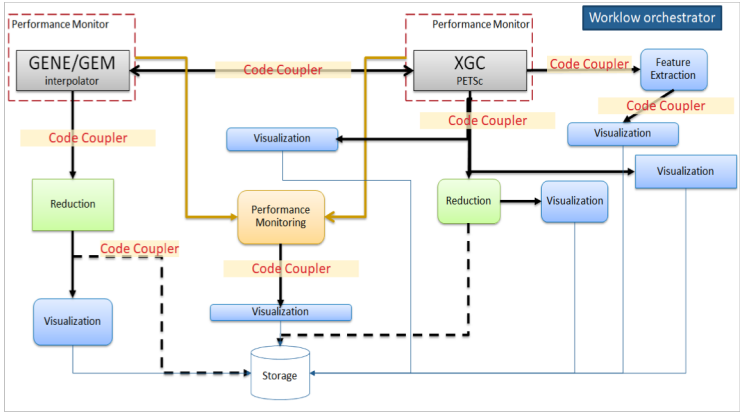
Enabling science understandability and reproducibility through rich data formats, metadata, and provenance

- **Acquiring, storing, analyzing, and maintaining science metadata** enabling human and machine access.
- **Organizing and managing the relationships** within and between science data to enable query and browsing.
- **Documenting the science data lifecycle**, from creation to preservation in support of reproducibility and verification.



Graph-based methods of organizing and interacting with metadata are one possible alternative to current approaches.

D. Dai et al, "GraphTrek: Asynchronous Graph Traversal for Property Graph Based Metadata Management," Cluster 2015, September 2015.



The schematic illustrates a complex set of processes to couple two fusion codes on an LCF, along with reducing, analyzing, and visualizing the results. Coupling can require temporarily storing data in another location in the system in order to free resources for computation.

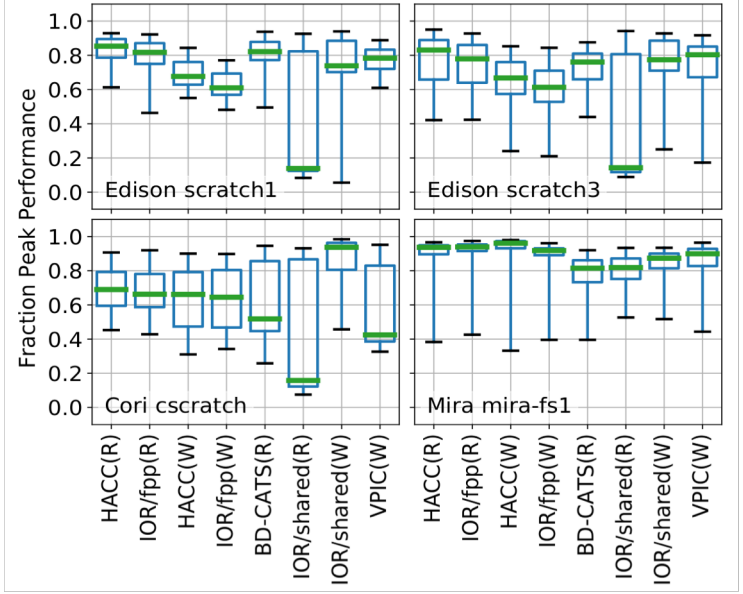
J. Y. Choi et al., "Coupling Exascale Multiphysics Applications: Methods and Lessons Learned," 14th IEEE International Conference on e-Science, 2018.

Accelerating scientific discovery through support of in situ and streaming data analysis

- Improving exploration by **exposing intent**
- Providing means for **multimodal analysis** by enabling one data source to serve multiple, different research efforts
- **Establishing common interfaces** for data stream access and processing
- Supporting different **reliability and performance requirements** for data streams

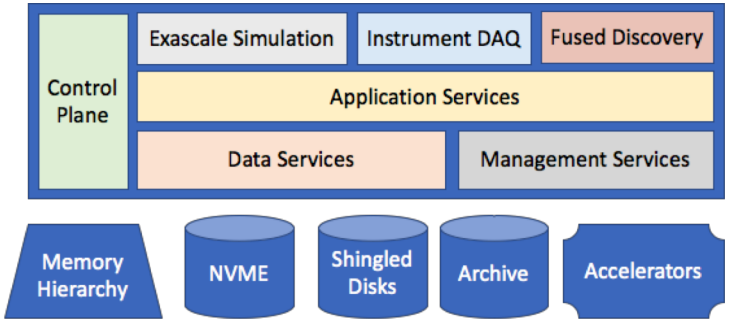
Enhancing SSIO usability, performance, and resilience through monitoring, prediction, and automation

- **Enabling real-time and post hoc analysis** through instrumentation, capture, and retention of monitoring data
- **Predicting behavior** through workload, software stack, and architectural modeling
- **Automatically adapting SSIO systems** in response to changes in their workload and/or environment.



Performance variance for 8 scientific I/O motifs, across 4 different parallel file systems, over a year of production activity. This figure illustrates performance variability due workload contention and and other hidden factors on large-scale storage systems

Lockwood et al., "A Year in the Life of a Parallel File System." SC'18, Nov. 2018.



In this diagram we show a set of building blocks for composed application storage services. One advantage of composed services is the improved efficiency available by addressing application requirements with a specific storage protocol and media type.

For example, a machine-learning parameter server may require NVME levels of performance for model updates, while a simulation post-processing step resulting in large analysis datasets may be streamed into disks for subsequent rendering.

Improving efficiency and integrity of data movement and storage through architecture of systems and services

- **Composing advanced storage services** that target specific DOE science workflows
- **Placing, moving, and locating data in the storage hierarchy** to meet application I/O workload needs and improve center-wide performance
- **Capitalizing on diverse media characteristics** to design efficient storage hierarchies


CONCLUDING REMARKS

NSF is interested, too!

- **Introspection and provenance**
  - Tracing and “demultiplexing” workloads
  - Correlating provenance from multiple layers
- **In situ and in transit data analysis**
  - Implications for storage designs
- **Confluence of AI and storage**
  - Storage in service of AI
  - AI in service of storage
- **New storage architectures**
  - Architectures for nonvolatile memory
  - Composition from building blocks

Data Storage Research Vision 2025

Report on NSF Visioning Workshop held May 30–June 1, 2018

George Amvrosiadis[†], Ali R. Butt[¶], Vasily Tarasov[‡], Erez Zadok[⋆], Ming Zhao[§]

Irfan Ahmad, Remzi H. Arpaci-Dusseau, Feng Chen, Yiran Chen, Yong Chen, Yue Cheng, Vijay Chidambaram, Dilma Da Silva, Angela Demke-Brown, Peter Desnoyers, Jason Flinn, Xubin He, Song Jiang, Geoff Kuenning, Min Li, Carlos Maltzahn, Ethan L. Miller, Kathryn Mohror, Raju Rangaswami, Narasimha Reddy, David Rosenthal, Ali Saman Tosun, Nisha Talagala, Peter Varman, Sudharshan Vazhkudai, Avani Waldani, Xiaodong Zhang, Yiying Zhang, and Mai Zheng.

[†]Carnegie Mellon University, [¶]Virginia Tech, [‡]IBM Research, [⋆]Stony Brook University, [§]Arizona State University

February 2019

Executive Summary

With the emergence of new computing paradigms (e.g., cloud and edge computing, big data, Internet of Things (IoT), deep learning, etc.) and new storage hardware (e.g., non-volatile memory (NVM), shingled-magnetic recording (SMR) disks, and kinetic drives, etc.), a number of open challenges and research issues need to be addressed to ensure sustained storage systems efficacy and performance. The wide variety of applications demand that the fundamental design of storage systems should be revisited to support application-specific and application-defined semantics. Existing standards and abstractions need to be reevaluated; new sustainable data representations need to be designed to support emerging applications. To take advantage of hardware advancements, new storage software designs are also necessary in order to maximize overall system efficiency and performance.

Therefore, there is a urgent need for a consolidated effort to identify and establish a vision for storage systems research and comprehensive techniques that provide practical solutions to the storage issues facing the information technology community. To address this need, the National Science Foundation’s (NSF) “Visioning Workshop on Data Storage Research 2025” brought together a number of storage researchers from academia, industry, national laboratories, and federal agencies to develop a collective vision for future storage research, as well as to prioritize near-term and long-term storage research and scientific investigations. In-depth discussions were carried out at the workshop along four major themes: (1) Storage for Cloud, Edge, and IoT Systems; (2) AI and Storage; (3) Rethinking Storage Systems Design; and (4) Evolution of Storage Systems with Emerging Hardware. The participants especially underscored the need for focused educational and training activities to instill storage system tools and technologies in the next generation of researchers and IT practitioners. Finally, the development of shared, scalable, and flexible community infrastructure to enable and sustain innovative storage research and verifiable evaluation was also discussed. This report presents the findings from these discussions.

1 Introduction

There are a number of open challenges and research issues that need to be addressed both in the short and long term to ensure sustained storage systems efficacy and performance.

1

https://dl.acm.org/citation.cfm?id=3316807

Storage and I/O are fun again!

High-Energy Physics Event Store (HEPnOS)

Goals

- Manage high-energy physics event data through multiple analysis phases
- Retain data in the system to accelerate analysis

Features

- Write-once, read-many
- Hierarchical namespace (datasets, runs, subruns)

Particle Trajectory Assembly (DeltaFS)

Goals

- Extreme scale file system metadata
- In-situ indexing with fast file retrieval

Features

- Specialized directories to efficiently support trillions of files in a single directory
- Software routing to scalably manage connections

Deep NN Model Cache (FlameStore)

Goals

- Store deep neural network models during a deep learning workflow
- Retain most promising candidate models

Features

- Flat namespace
- Python API (Keras models)

In-System Object Store (Mobject)

Goals

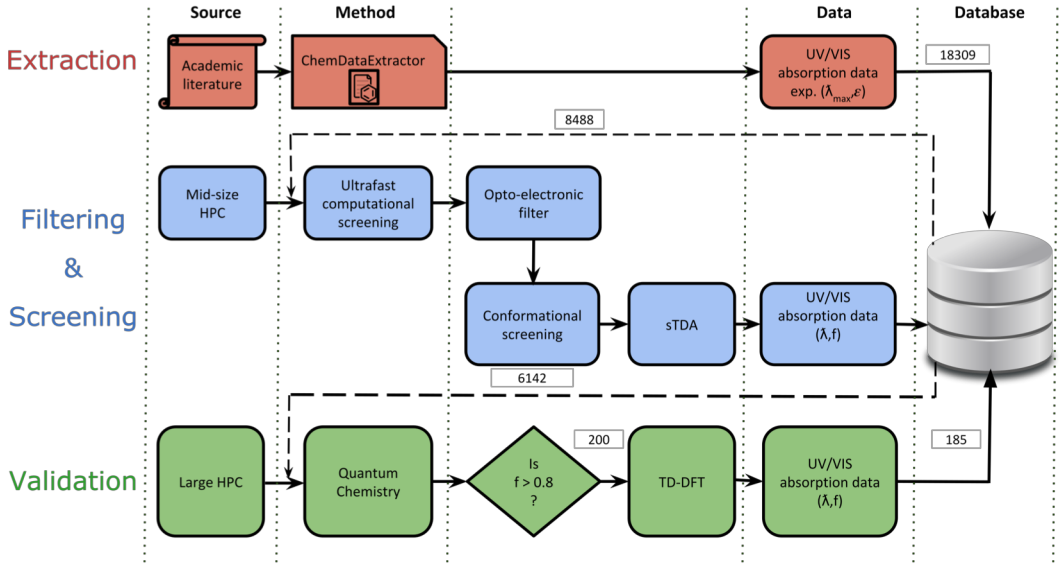
- Provide familiar model as alternative to POSIX

Features

- Concurrent read/write
- Flat namespace
- RADOS client API (subset)

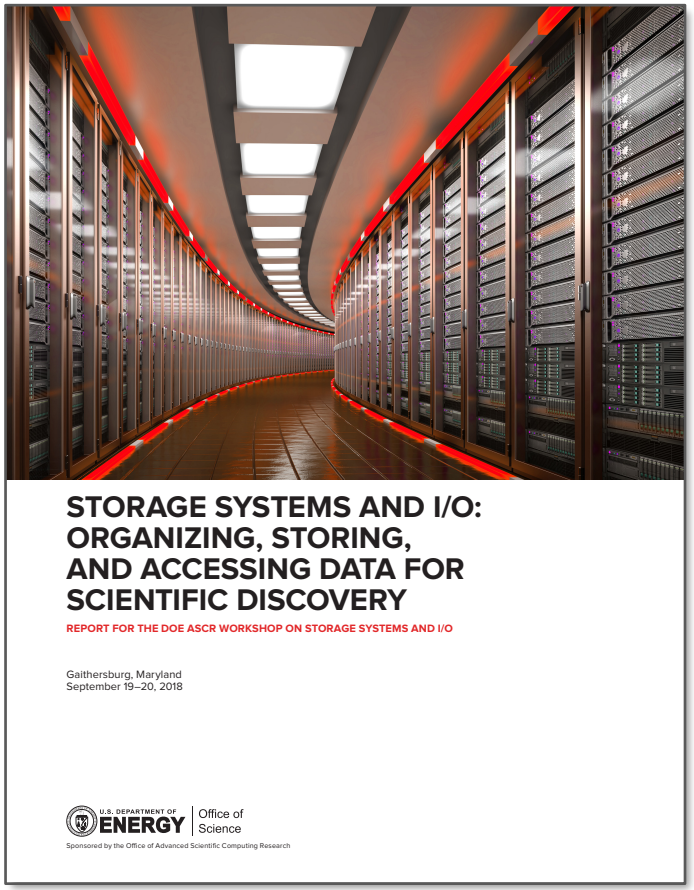
https://www.mcs.anl.gov/research/projects/mochi/

Tackling modern science challenges on future platforms requires solutions from many SSIO “sub-disciplines”



ASCR excels at bringing these sorts of multi-disciplinary teams together.

Thanks! Questions?



Soon!


ADDITIONAL MATERIALS

Attendees

Mark Ainsworth, Brown University
George Amvrosiadis, Carnegie Mellon
Michael Bender, Stony Brook University
John Bent, DataDirect Networks
Wes Bethel, LBNL
Laura Biven, DOE
David Bonnie, LANL
Ron Brightwell, SNL
Benjamin Brown, DOE
Ali Butt, Virginia Tech
Suren Byna, LBNL
Raghunath Chandrasekar, Amazon Web Services
Antonio Cortes, Barcelona Supercomputing Center
Matthew Curry, SNL
Ewa Deelman, Univ. of Southern California
Andreas Dilger, Whamcloud
Daniel Ernst, Cray, Inc.
Lance Evans, Cray, Inc.
Evan Felix, PNNL
Greg Ganger, Carnegie Mellon University
Ada Gavrilovska, Georgia Tech
Elsa Gonsiorowski, LLNL
Gary Grider, LANL
Kevin Harms, ANL
Graham Heyes, Jefferson Lab
Dean Hildebrand, Google
Terry Jones, ORNL
Kristy Kallback-Rose, NERSC/LBNL
Dimitrios Katramatos, BNL
Scott Klasky, UT-Battelle, ORNL
Quincey Koziol, LBNL
Lingda Li, BNL
Glenn K. Lockwood, LBNL
Jay Lofstead, SNL
Johann Lombardi, Intel Corporation
Darrell Long, UC Santa Cruz
Xiaosong Ma, Qatar Computing Research Institute
Carl Maltzahn, UC Santa Cruz
Kathryn Mohror, LLNL
Thomas Ndousse-Fetter, DOE
Bogdan Nicolae, ANL
Lucy Nowell, DOE
Manish Parashar, Rutgers University
Amedeo Perazzo, SLAC
Tom Peterka, ANL
Robinson Pino, DOE
Eric Pouyoul, SND/ESnet LBNL
Lavanya Ramakrishnan, LBNL
Robert Ross, ANL
Sonia Sachs, DOE
Joel Saltz, Stony Brook University
Malachi Schram, PNNL
Bradley Settlemyer, LANL
Michela Taufer, UT Knoxville
Deneise Terry, ORISE
Angie Thevenot, DOE
Sudharshan Vazhkudai, ORNL
Lee Ward, SNL
Jack Wells, ORNL
Matthew Wolf, ORNL
Weikuan Yu, Florida State University

Workshop Agenda

Wednesday	
Time	**Activity**
8:15am – 8:35am	Welcome (Barb Helland) and opening remarks (Lucy Nowell) Reminder of charge, overview of meeting, safety, etc.
8:35am – 8:50am	Talk: Experimental and Observational Data Wes Bethel
8:50am – 9:05am	Talk: Streaming Data Graham Heyes
9:05am – 9:20am	Talk: Workflow Management Tom Peterka
9:20am – 10:00am	Talk: Science requirements for SSIO at the LCFs Jack Wells & Kevin Harms
10:00am – 10:30am	Break
10:30am – 11:45am	Panel: Applications and Facilities Requirements (Application Pull) Moderator: Kathryn Mohror Participants: Wes Bethel, Graham Heyes, Jack Wells, Kevin Harms, Evan Felix, Tom Peterka, Kristy Kallback-Rose
11:45am –12:35am	Lunch
12:35pm – 2:05pm	Working Session 1: Integrating with Science Workflows Moderator: Scott Klasky Scribe: Brad Settlemyer
2:05pm – 2:35pm	Break
2:35pm – 4:05pm	Working Session 2: Understanding SSIO Systems Moderator: Rob Ross Scribe: Galen Shipman
4:05pm – 4:25pm	Break
4:25pm – 5:55pm	Working Session 3: Streaming Data Moderator: Matt Wolf Scribe: Glenn Lockwood

Thursday	
Time	**Activity**
8:15am – 8:30am	Talk: Extreme Heterogeneity Workshop Report Lucy Nowell
8:30am – 8:50am	Talk: Storage Technologies Gary Grider
8:50am – 9:10am	Talk: Memory Technologies; Blurring the Lines Dan Ernst
9:10am – 10:15am	Panel: Storage Technologies (Tech Push Panel) Moderator: Lee Ward Participants: Gary Grider, Kevin Harms, Eric Pouyoul, Dan Ernst, Lance Evans
10:15am – 10:45am	Break
10:45am – 12:15pm	Working Session 4: Heterogeneous/multi-tier storage systems Moderator: Kathryn Mohror Scribe: Kevin Harms
12:15pm – 1:15pm	Lunch
1:15pm – 1:30pm	Talk: ISDM Workshop Tom Peterka
1:30pm – 3:00pm	Working Session 5: Metadata, Name Spaces, and Provenance Moderator: Lee Ward Scribe: Quincey Koziol
3:00pm –- 3:25pm	Break
3:25pm – 4:55pm	Working Session 6: HW/SW architectures Moderator: Brad Settlemyer Scribe: Rob Ross
4:55pm – 5:00pm	Closing remarks and adjourn