DOE Storage Systems and Input/Output (SSIO) Workshops

Held December 8-11, 2014

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

Organizers

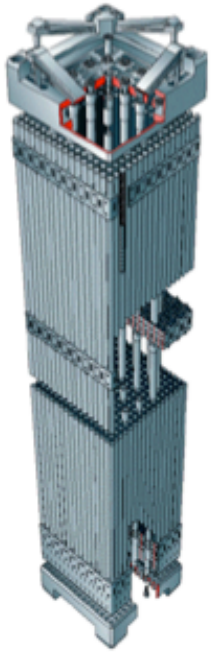
- Rob Ross (ANL)
- Gary Grider (LANL)
- Evan Felix (PNNL)
- Mark Gary (LLNL)
- Scott Klasky (ORNL)
- Ron Oldfield (SNL)
- Galen Shipman (LANL)
- John Wu (LBNL)

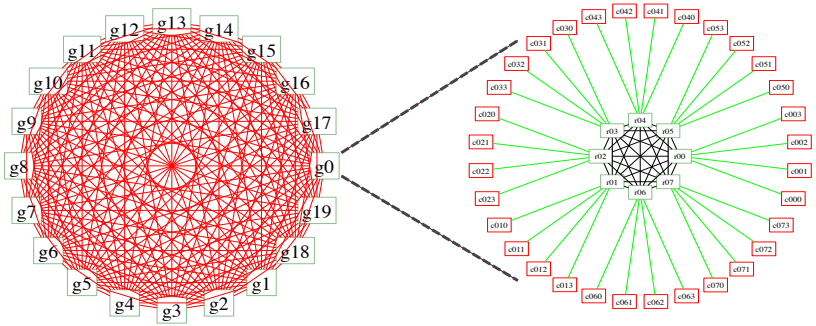
Sponsor

- Lucy Nowell (ASCR)

What is Storage Systems and Input/Output (SSIO)?

- Everything from the low-level parallel file system and archival storage up to libraries that serve as the interfaces to applications
- New challenges mandate new R&D:
  - Deeper storage hierarchies
  - Increasing scale(s), complex topologies
  - Demand for greater resilience
  - New science workflows



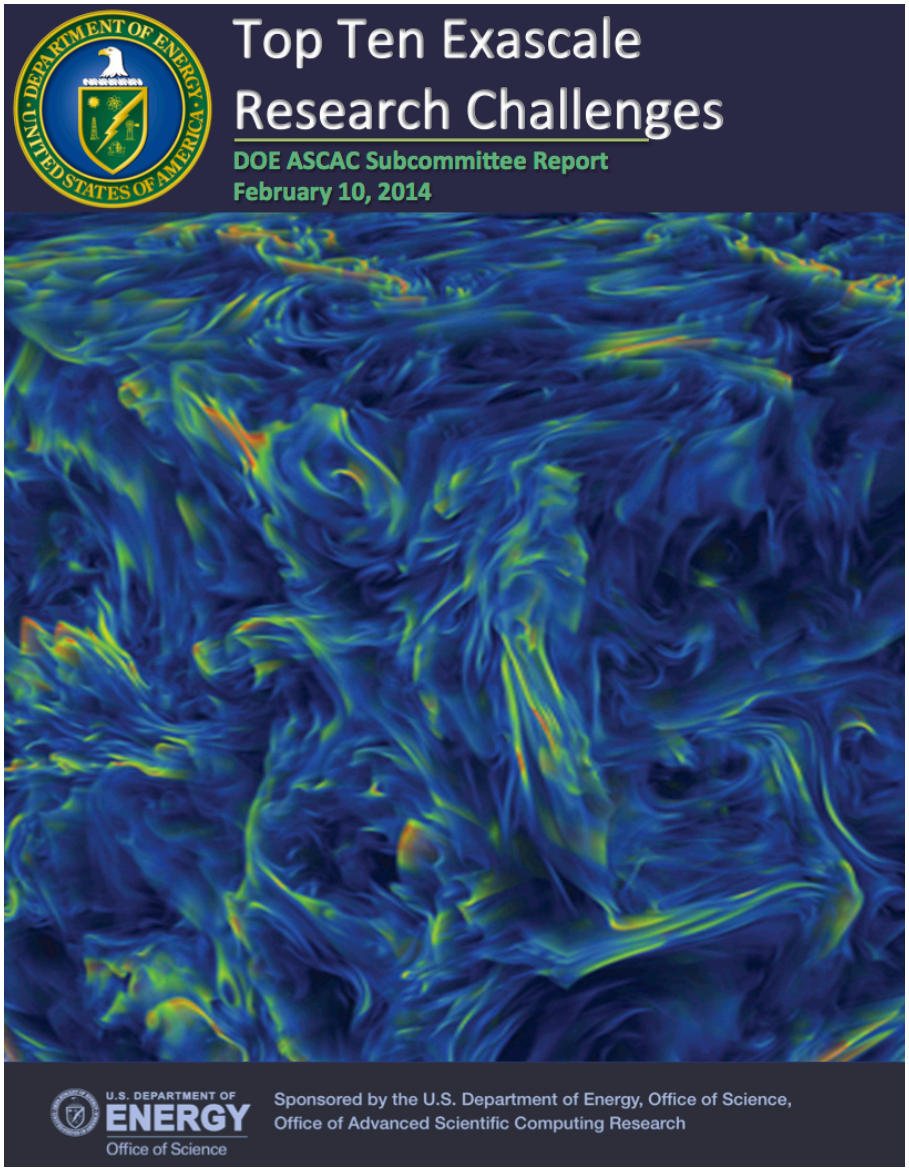


Storage, I/O, and Exascale

#6 Data management: Creating data management software that can handle the volume, velocity and diversity of data that is anticipated.

But also:

#4 Scalable System Software: Developing scalable system software that is power- and resilience- aware.



Goals and Process

- Goals:
  - Review SSIO requirements for simulation-driven activities
  - Assess state of the art
  - Identify priority research directions in SSIO

- 3 Workshops
  - Requirements gathering
  - Cross-cutting computer science
  - The SSIO meeting itself

Organizers
- Rob Ross (ANL)
- Gary Grider (LANL)
- Evan Felix (PNNL)
- Mark Gary (LLNL)
- Scott Klasky (ORNL)
- Ron Oldfield (SNL)
- Galen Shipman (LANL)
- John Wu (LBNL)

Sponsor
- Lucy Nowell (ASCR)

Workshop 1: Requirements Gathering

Goal: Define a set of use cases for future systems to inform SSIO R&D community.

- Application representatives presented future requirements
- Answered a detailed set of questions on topics such as:
  - Composition of jobs, phases of I/O
  - Structure of data, methods of query
  - Methods of data reduction during runtime (i.e., *in situ*)
  - Expected uses of NVRAM in future systems
  - Archival storage use, provenance capture

Science/Mission Reps

- Salman Habib (HEP)
- Rob Neely (ASC)
- Varis Carey (ExaCT)
- David Rogers (ASC)
- Dave Richards (ExMatEx)
- Michael Glass (ASC)

Workshop 2: Cross-cutting Computer Science

Goal: Identify critical SSIO requirements and points for coordination between SSIO and other CS areas.

- Experts in related CS areas presented their views on how their area intersects with SSIO
  - Operating systems
  - Networking
  - Workflow
  - Resilience
  - Analysis and visualization
  - Collaborative technologies

Crosscutting CS Representatives

- Pete Beckman (OS/Runtime)
- Kerstin Kleese van Dam (Workflow)
- Ian Foster (Collab. Technologies)
- Oliver Rubel (Analysis and Vis.)
- Nathan Debardeleben (Resilience)
- Maya Gokhale (Analytics)
- Jay Lofstead (LAN Networking and OS)

Workshop 2: Cross-cutting Computer Science

In situ Infrastructure

Dependencies and Relationships with SSIO (1/2)

In situ Data Analysis and Visualization I/O	Traditional Simulation I/O
Read data into the simulation: e.g, analyses across time, compare with observation etc.	Write only
Save complex analysis results: including surfaces, graphs, sparse matrices etc.	Write structured fields
More frequent "smaller" writes: Save reduced data and analysis results at higher temporal resolution **Irregular temporal intervals between writes:** Save data and analysis results when specific features/events are discovered	Write data at regular and often sparse time intervals
Potentially unbalanced I/O load: E.g., the data partitioning may be optimized for the simulation not DAV.	Simulations often optimized for load balancing

Slide from O. Rubel.

- Jay Lofstead (LAN Networking and OS)

Workshop 3: Identifying Research Directions

Goal: Identify potential research directions in SSIO for extreme scale DOE science.

- Initial talks summarized findings from prior workshops and other recent activities in the area
- Single track, open discussion, organized around five areas:
  - HW/SW architectures for SSIO
  - Metadata, name spaces, and provenance
  - Supporting science data
  - Integration with external services
  - Understanding SSIO systems

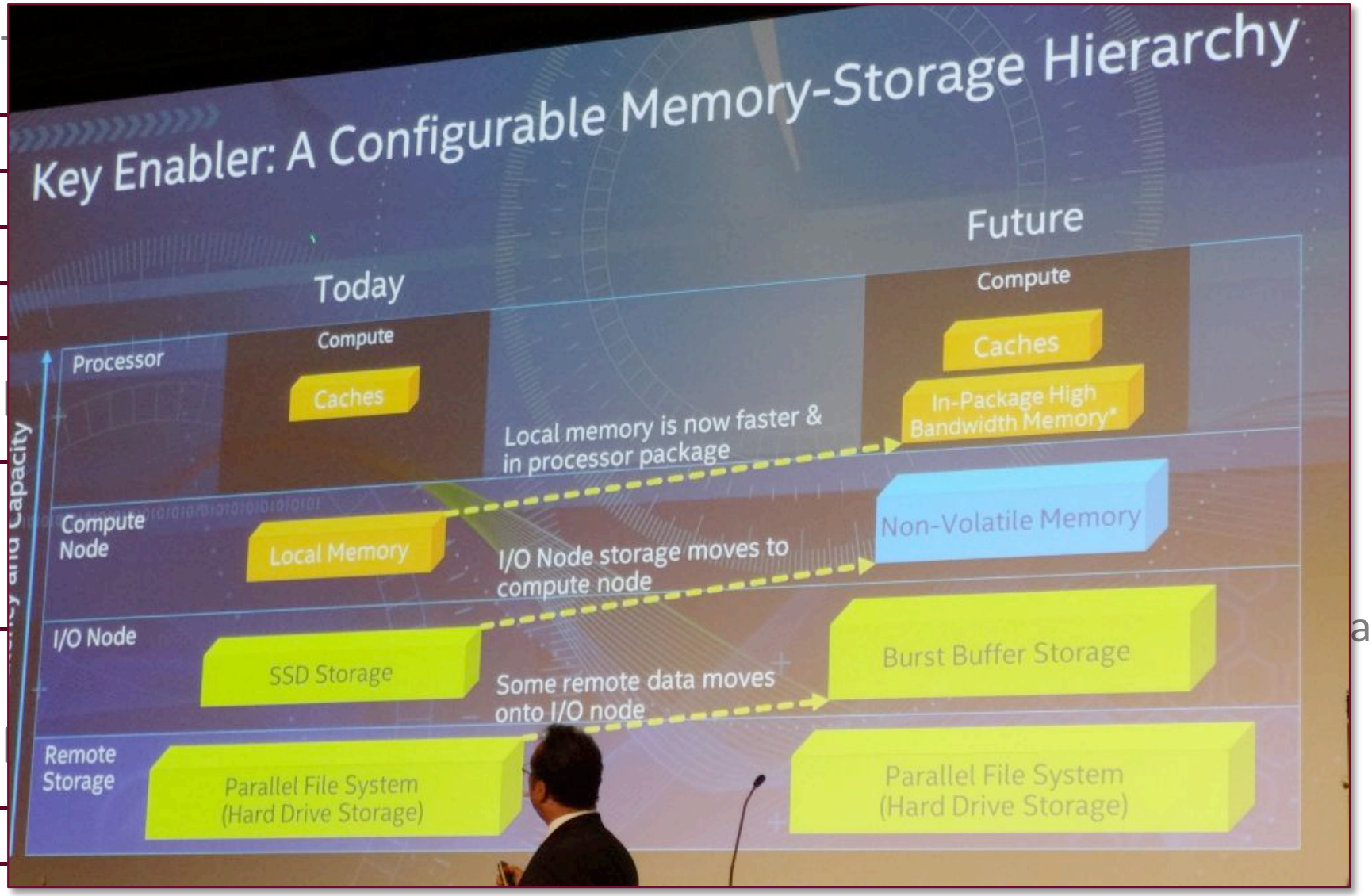
Workshop 3 Computer Scientist Participants

- Hasan Abbasi (ORNL)
- Eric Barton (Intel)
- Michael Bender (SUNY SB)
- John Bent (EMC)
- Suren Byna (LBNL)
- Phil Carns (ANL)
- John Chandy (UConn)
- Matt Curry (SNL)
- Bronis de Supinski (LLNL)
- Garth Gibson (CMU)
- Kevin Harms (ANL)
- Quincey Koziol (HDF Group)
- Bradley Kuszmaul (MIT)
- Wei-keng Liao (Northwestern)
- Darrell Long (UCSC)
- Carlos Maltzahn (UCSC)
- Meghan McClelland (Seagate)
- Ethan Miller (UCSC)
- Adam Moody (LLNL)
- Paul Nowoczynski (DDN)
- Manish Parashar (Rutgers)
- Narasimha Reddy (TAMU)
- Brad Settlemyer (LANL)
- Rajeev Thakur (ANL)
- Sudharshan Vazhkudai (ORNL)
- Lee Ward (SNL)
- Brent Welch (Google)
- Matt Wolf (GA Tech)
- Cornell Wright (LANL)
- Wenji Wu (FNAL)
- Erez Zadok (SUNY SB)

Hardware/Software Architectures

- Topics included
  - Network technologies and topologies
  - Solid-state storage in and near the HPC system
  - Compute-in-storage
  - System noise, reliability
  - Autonomics
- Findings
  - **Storage hierarchy is increasing in complexity**. Current organization methods (e.g., parallel file systems, archival management) must significantly change or be replaced to address this complexity.
  - **Scientists need an integrated view of storage resources.** New metadata capabilities and integration with external storage are also needed.
- Priorities for research
  - Managing deep and heterogeneous storage hierarchies
  - Alternative management paradigms to the file system model

Hardware/Software Architectures

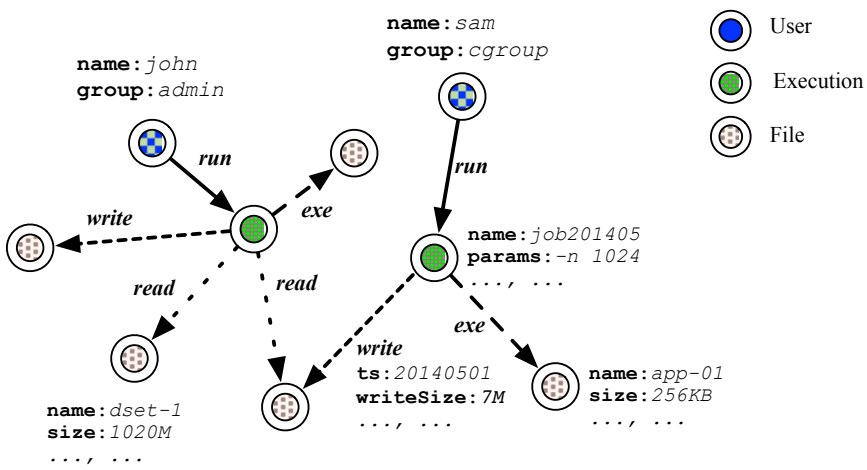


ata

Raj Hazra, ISC, July 2015. Image from N. Hemsworth, “One Single System Architecture to Rule Them All,” July 20, 2105.

Metadata, Name Spaces, and Provenance

- Topics included
  - Metadata and alternative data stores
  - Automating provenance capture, connection to other services
- Finding
  - **New requirements for validation of results will change role of metadata in DOE applications.** New methods for capturing provenance and exploring datasets will be needed.
- Priorities for research
  - New methods of management of rich metadata
  - Breaking away from the current file model

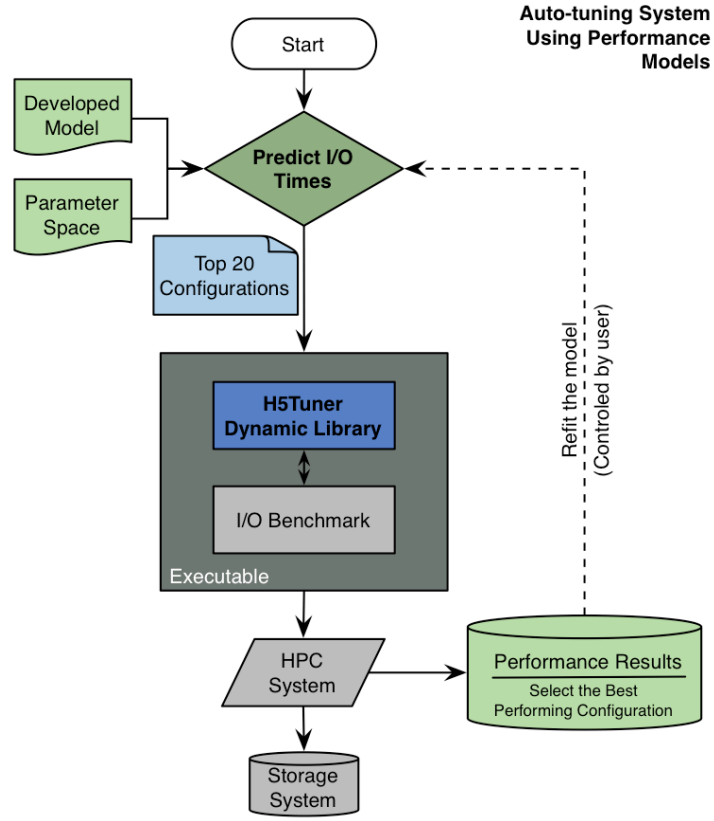


Graph-based methods of organizing and interacting with metadata are one possible alternative to current approaches.

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

Supporting Science Data

- Topics included
  - Programming model integration
  - SSIO services in support of workflow
  - Self-tuning libraries
  - Data abstractions
- Findings
  - **Scientist productivity is tied to ability to represent and interact with complex and specialized data.**
  - **Alternative programming languages and increased need for workflow support drive new SSIO research.**
- Priorities for research
  - New generation of I/O middleware and <u>services</u> to support new programming abstractions and workflows



Autotuning I/O middleware can provide dramatically improved performance with limited search time over possible configurations.

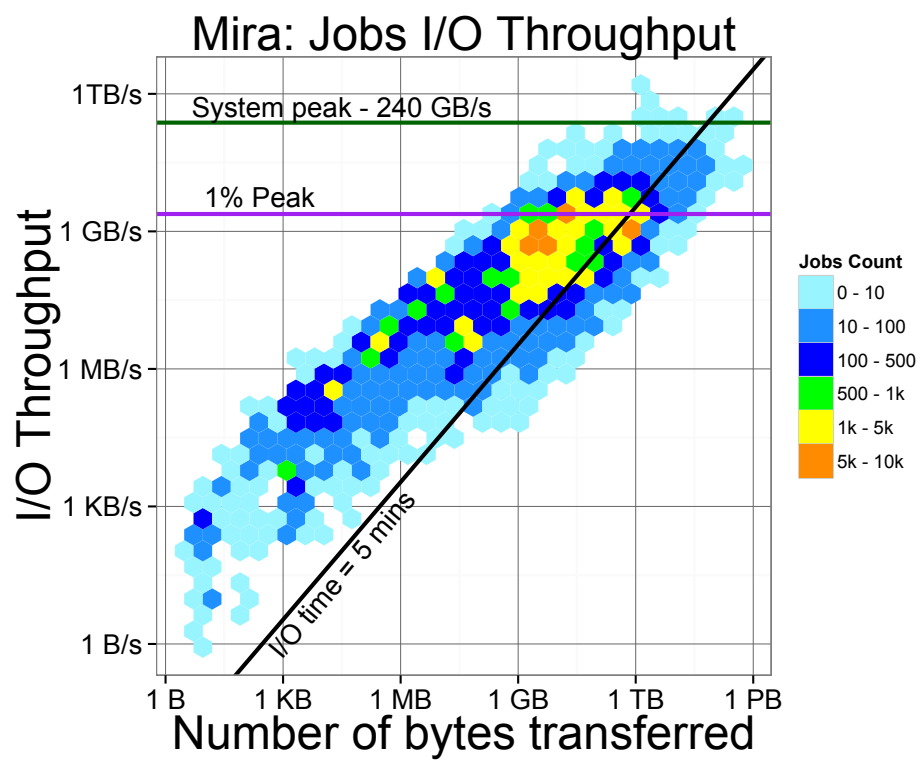
B Behzad et al, "Improving Parallel I/O Autotuning with Performance Modeling," HPDC 2014, June 2014.

Integration with External Services

- Topics included
  - Scheduling and resource management
  - System monitoring
  - Workflow and orchestration
  - Archival storage
- Finding
  - **Current SSIO designs hindered by isolation from system-level resource management, monitoring, and workflow systems.**
- Not a research direction in itself, but rather influences direction of research in other areas

Understanding Storage Systems and I/O

- Topics included
  - Workload characterization
  - Modeling and simulation
  - Designing for understandability
- Findings
  - **Many important aspects of application and system behavior are obscured from view.** Better understanding is needed to maximize SSIO effectiveness.
- Priorities for research
  - Improve our ability to characterize storage activities to model and predict the behavior of SSIO activities on future systems.



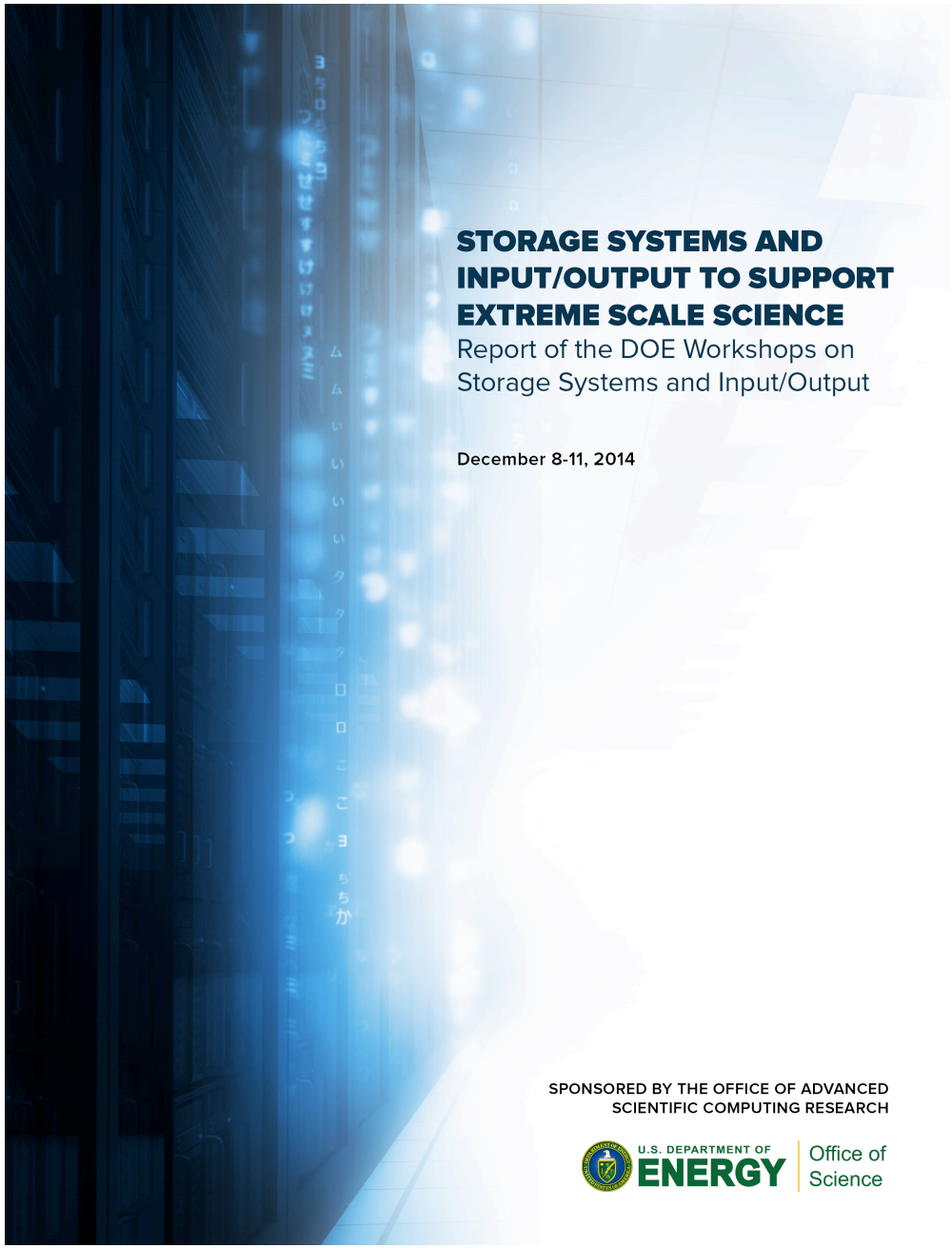
Characterization of I/O traffic on production systems helps identify opportunities for improvement, leading to more productive use of DOE systems.

H. Luu et al, "A Multiplatform Study of I/O Behavior on Petascale Supercomputers," HPDC 2015, June 2015.

For more information…

Thanks to:

- Gary Grider (LANL)
- Evan Felix (PNNL)
- Mark Gary (LLNL)
- Scott Klasky (ORNL)
- Ron Oldfield (SNL)
- Galen Shipman (LANL)
- John Wu (LBNL)
- Lucy Nowell (ASCR)



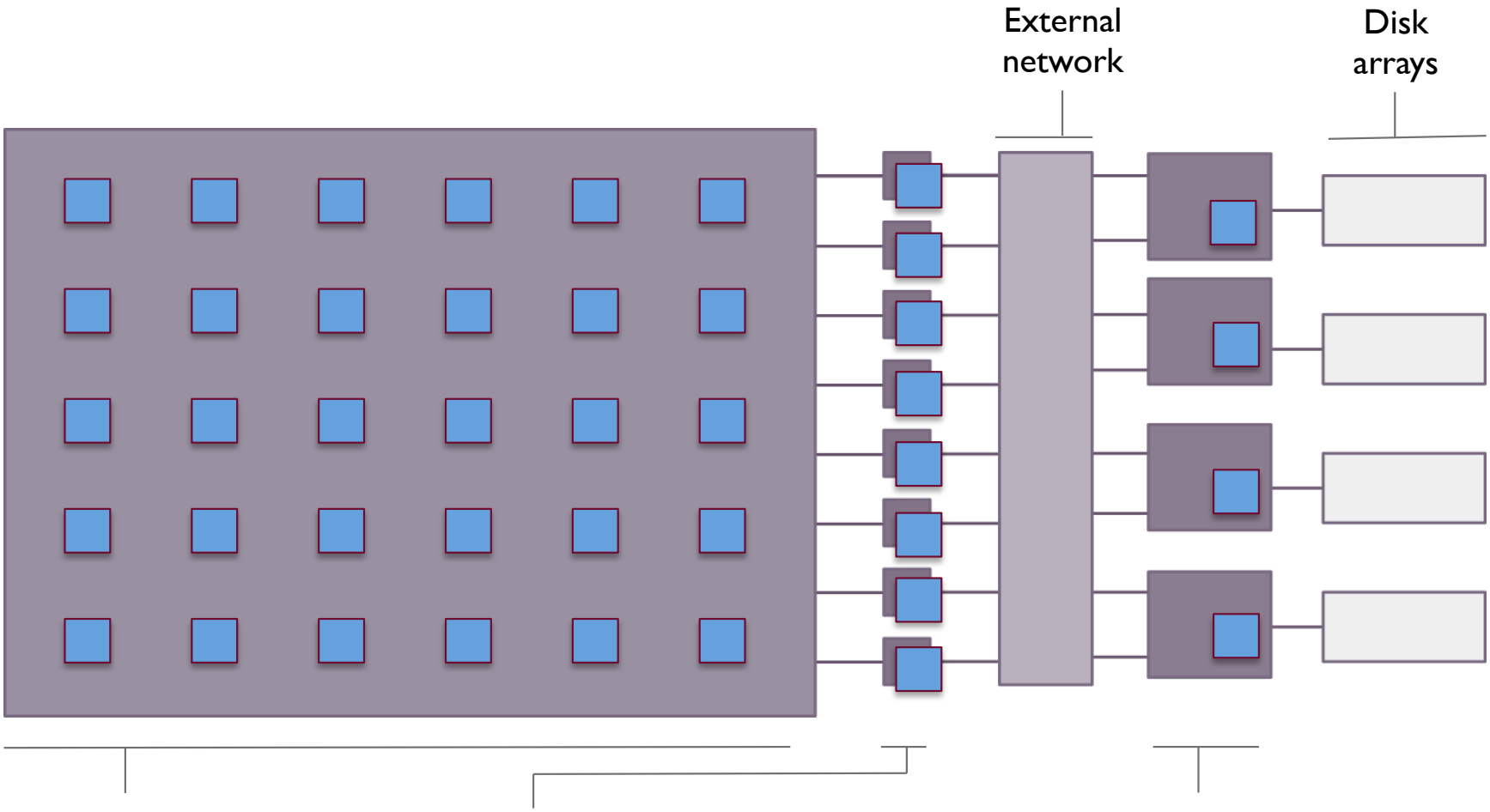
http://science.energy.gov/~/media/ascr/pdf/programdocuments/docs/ssio-report-2015.pdf

15-1338 Storage and I/O for Extreme Scale Science

- Themes:
  - Measurement and Understanding
  - Scalable Storage Software Infrastructure
  - New Paradigms in SSIO

- Proposals were due July 13, 2015

Ongoing Burst Buffer Discussion



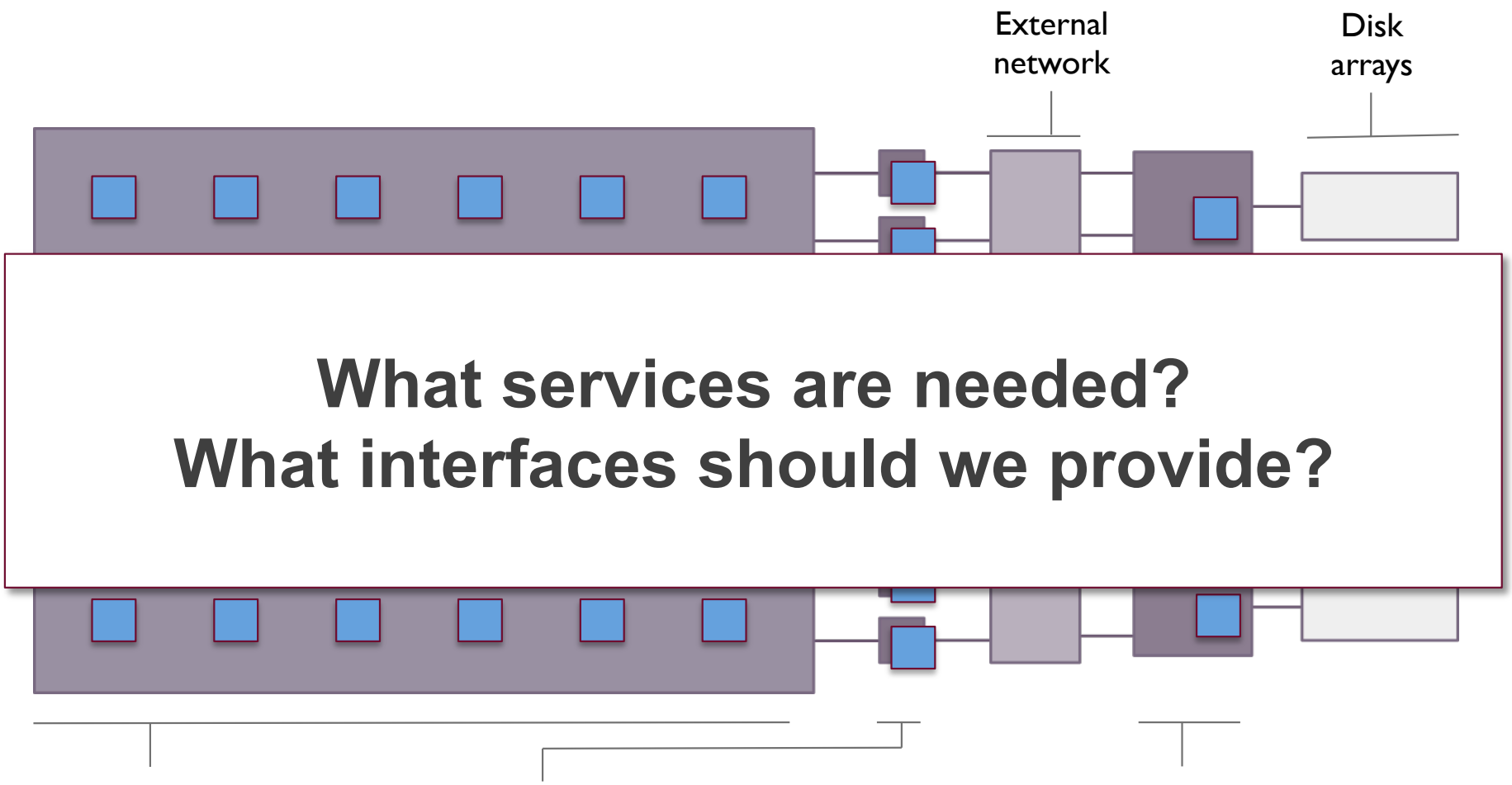
Compute nodes run application processes.

I/O forwarding nodes (or I/O gateways) shuffle data between compute nodes and external resources, including storage.

Storage nodes run the parallel file system (or alternative).

- Possible solid-state storage (burst buffer) location

Ongoing Burst Buffer Discussion

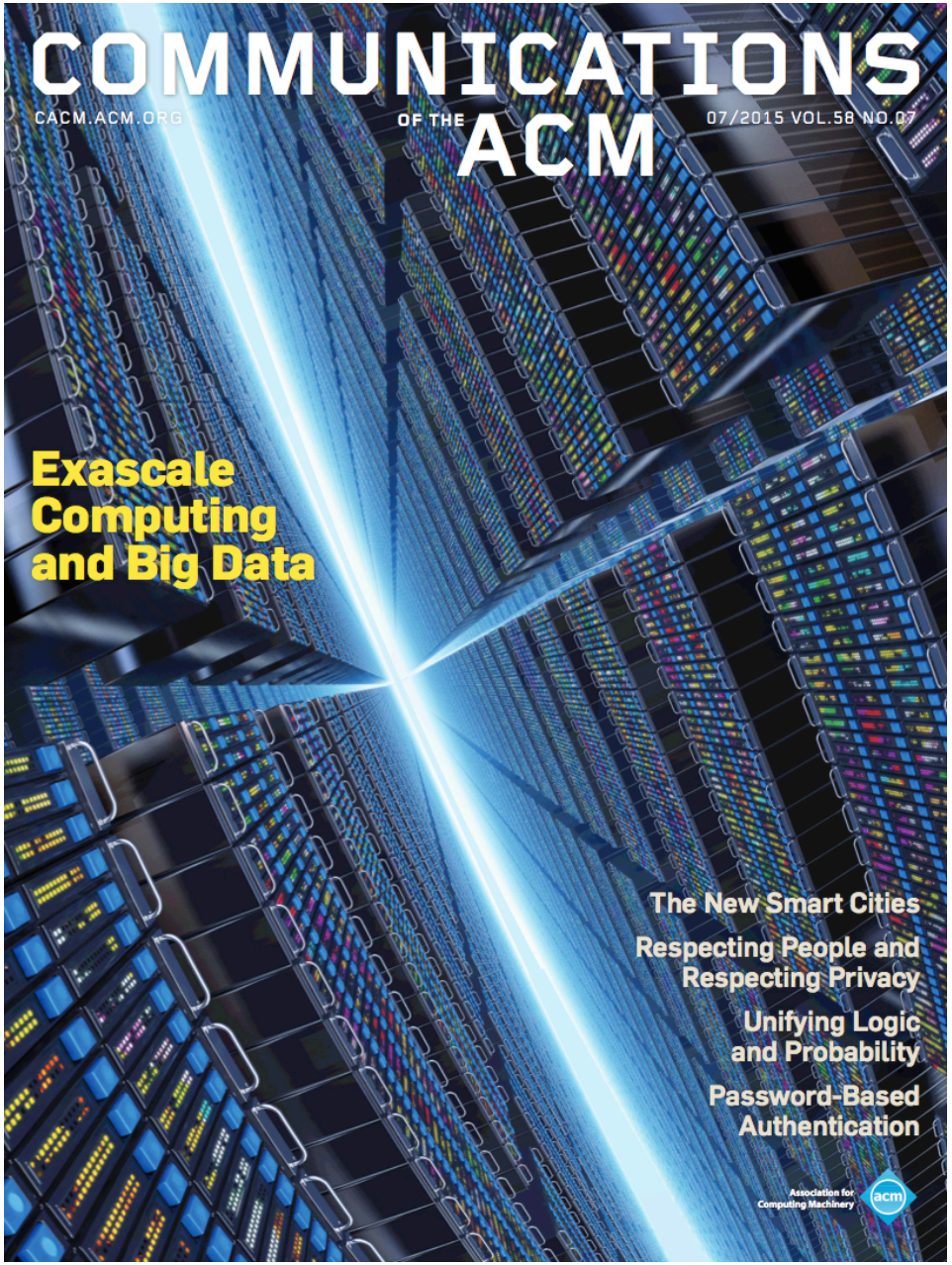


What services are needed?
What interfaces should we provide?

Compute nodes run application processes.

I/O forwarding nodes (or I/O gateways) shuffle data between compute nodes and external resources, including storage.

Storage nodes run the parallel file system (or alternative).

Looking Forward

- Experimental and observational data brings new challenges
- HPC has a role, SSIO also has a role…