ASCR WORKSHOP ON IN SITU DATA MANAGEMENT

Enabling Scientific Discovery
from Diverse Data Sources

January 28–29, 2019

U.S. DEPARTMENT OF ENERGY

The Scientific Impact of In Situ Data Management

In January 2019, the U.S. Department of Energy (DOE), Office of Science (SC) program in Advanced Scientific Computing Research (ASCR) convened a workshop to identify priority research directions (PRDs) for **In Situ Data Management (ISDM).**

The workshop defined ISDM as the practices, capabilities, and procedures to control the organization of data and enable the coordination and communication among heterogeneous tasks, executing simultaneously in a high-performance computing (HPC) system, cooperating toward a common objective.

A fundamental finding of this workshop is that the methodologies used to manage data among a variety of tasks in situ can be used to facilitate scientific discovery from a variety of different data sources—simulation, experiment, and sensors, for example—and that being able to do this at a variety of computing scales will benefit real-time decision making, design optimization, and data-driven scientific discovery across the SC mission space. Applications wanting to make use of the in situ capabilities include those where data analysis feeds back to the simulation, decisions are made autonomously, big data or machine learning is among the tasks to be coordinated, and computations need to be completed in real time.

The workshop addressed the seven topic areas shown in Figure 1. Discussion sessions covered each area, and workshop participants were asked to identify key challenges and opportunities, list the potential benefits to the DOE-ASCR mission of addressing those challenges, and synthesize candidate research directions for each topic. Those research directions were further distilled to a set of final priority research directions (PRDs).

A summary of the workshop PRDs appears on the following pages. The full workshop report, with more details of the PRDs and summaries of each discussion session, will be available at DOI: 10.2172/1493245.

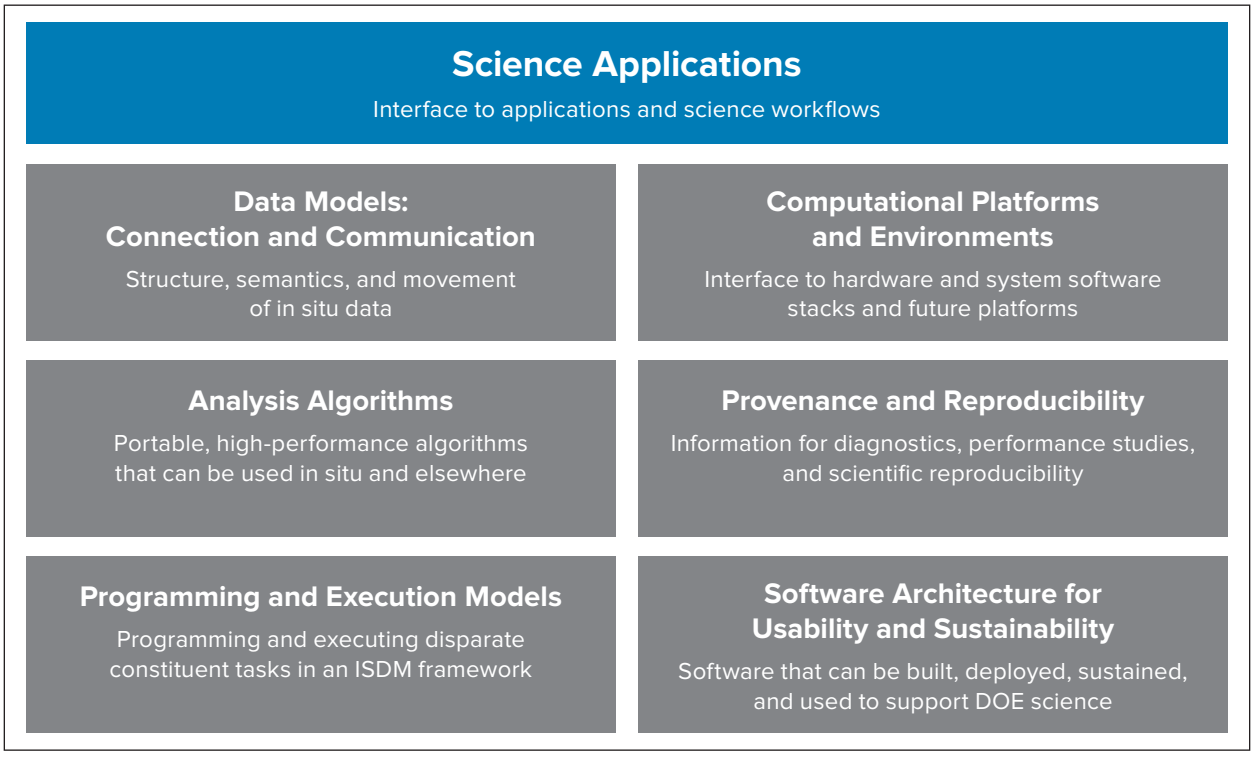
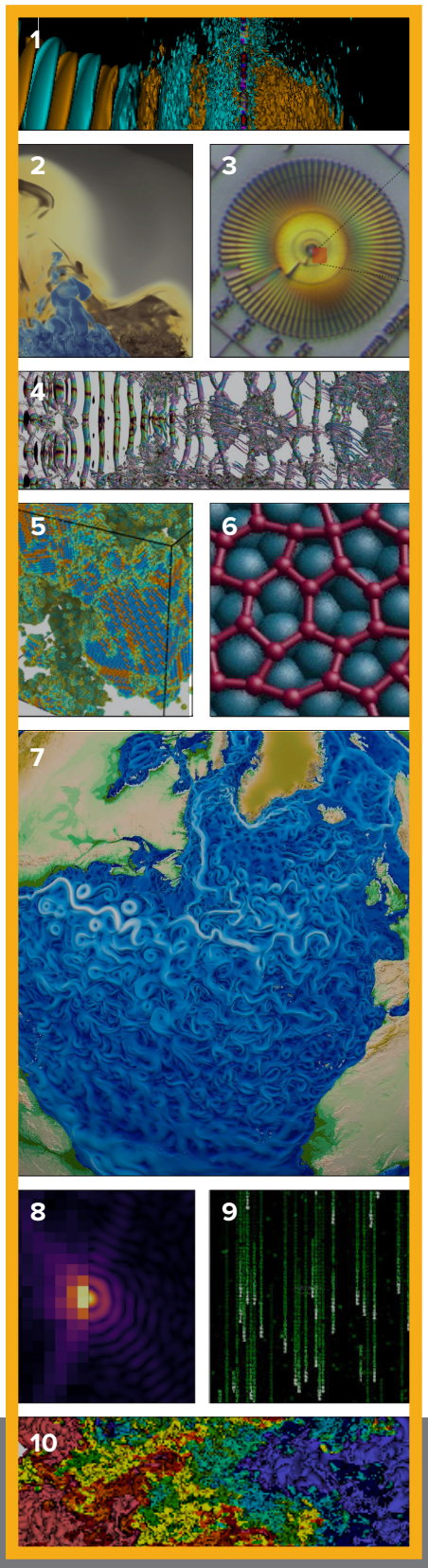


Figure 1: Workshop topic areas.

Priority Research Directions

Pervasive ISDM: Apply ISDM methodologies and in situ workflows at a variety of platforms and scales.	**In Situ Algorithms:** Redesign data analysis algorithms for the in situ paradigm.	**Composable ISDM:** Develop interoperable ISDM components and capabilities for an agile and sustainable programming paradigm.
Key questions: *How can ISDM methodologies help meet the needs for real-time, high-velocity data applications at the edge and other non-HPC platforms? How can ISDM enable science at experimental and observational facilities?* A changing landscape of use cases is driving new applications of ISDM. The ability to execute the same ISDM tasks and workflows across a spectrum of computational platforms, spanning high-performance supercomputers to experimental detectors and even embedded devices, will reduce human effort and improve portability by applying consistent computing methods.	**Key questions:** *How should in situ algorithms be designed to make most of the available resources? What new classes of data transformations can profit from in situ data access in the presence of constraints imposed by other tasks?* The in situ environment for data processing and analysis differs substantially from the post hoc environment, requiring fundamentally new algorithms and approaches. Progress will benefit from multidisciplinary approaches that holistically consider the opportunities, constraints, and user needs of in situ analysis.	**Key questions:** *Can the composition of ISDM software components maximize programmer productivity and usability? What design decisions of ISDM software components promote their interoperability in order to ensure the long-term utility of ISDM software for the science community?* The flexible composition of interoperable ISDM software components will enable developers and end-users to choose from an array of widely available tools, thereby increasing productivity, portability, and usability, and will ultimately result in agile and reusable software.
Co-designed ISDM: Coordinate the development of ISDM with the underlying system software so that it is part of the software stack.	**Controllable ISDM:** Understand the design space of autonomous decision-making and control of in situ workflows.	**Transparent ISDM:** Increase confidence in reproducible science, deliver repeatable performance, and discover new data features through the provenance of ISDM.
Key questions: *What abstractions, assumptions, and dependencies on system services are needed by ISDM? What information must be exchanged between the ISDM tools and the rest of the computing software stack to maximize performance and efficiency?* Understanding the interlayer dependencies so that ISDM becomes part of the software stack can facilitate connections between software layers, communicate semantic meaning, and realize efficient performance in HPC and other software stacks.	**Key questions:** *What metrics best describe the ISDM design space? How can that space be defined, codified, and evaluated to support design decision-making and control?* Understanding the space of ISDM parameters is crucial to making intelligent design decisions, both by humans and autonomously. The capability to optimize a constrained ISDM design space will enable predictable performance and scientific validity. Design metrics will promote knowledge sharing across communities.	**Key questions:** *How can provenance and metadata support data discoverability, reuse, and reproducibility of results? How can these artifacts be captured automatically and analyzed in situ, at the scale of DOE science?* In situ provenance and metadata are crucial to understanding scientific results, assessing correctness, and connecting underlying models and algorithms with workflow execution. The ability to capture and query provenance and metadata at scale and in situ will enable many diverse science needs.



Summary

Scientific computing will increasingly incorporate a number of different tasks that need to be managed along with the main simulation or experimental tasks—ensemble analysis, data-driven science, artificial intelligence, machine learning, surrogate modeling, and graph analytics—all nontraditional applications unheard of in HPC just a few years ago. Many of these tasks will need to execute concurrently, that is, in situ, with simulations and experiments sharing the same computing resources.

The workshop revealed two primary, interdependent motivations for processing and managing data in situ. The first motivation is that the in situ methodology enables scientific discovery from a broad range of data sources—HPC simulations, experiments, scientific instruments, and sensor networks—over a wide scale of computing platforms: leadership-class HPC, clusters, clouds, workstations, and embedded devices at the edge. The successful development of ISDM capabilities will benefit real-time decision making, design optimization, and data-driven scientific discovery. The second motivation is the need to decrease data volumes. ISDM can make critical contributions to managing large data volumes from computations and experiments to minimize data movement, save storage space, and boost resource efficiency—often while simultaneously increasing scientific precision.

The workshop identified six PRDs that highlight the components and capabilities needed for ISDM to be successful for the wide variety of applications discussed: making ISDM capabilities more pervasive, controllable, composable, and transparent, with a focus on greater coordination with the software stack, and a diversity of fundamentally new data algorithms.

Image credits:

1. Ion accelerator simulation. Image courtesy G. Weber, O. Ruebel, B. Loring, J.-L. Vay.
2. Asteroid impact simulation. Image courtesy J. Patchett, F. Samsel, G. Gisler.
3. High-energy X-ray test specimen. Image courtesy Y. Nashed.
4. Isosurfaces in direct numerical simulation of turbulent mixing. Image courtesy E. Duque.
5. Nucleated crystals in molecular dynamics simulation. Image courtesy O. Yildiz.
6. Silicene formation in molecular modeling. Image courtesy W. Usher, V. Pascucci, J. Insley, N. Ferrier, M. Papka, S. Rizzi, V. Vishwanath, J. Amstutz, I. Wald.
7. Ocean eddy formation. Image courtesy P. Wolfram, M. Peterson, T. Ringler.
8. Supersampled high-energy coherent diffraction. Image courtesy Y. Nashed.
9. Phenotype vectors in genome wide association studies. Image courtesy D. Jacobson.
10. Rayleigh-Taylor instability. Image courtesy B. Loring.

U.S. DEPARTMENT OF ENERGY | Office of Science