Priority Research Directions for In Situ Data Management

Enabling Scientific Discovery from Diverse Data Sources

“The future has a way of arriving unannounced.”
– George Will
Introduction

“The secret of getting ahead is getting started.”

– Mark Twain
Definition of In Situ Data Management (ISDM)

The practices, capabilities, and procedures to control the organization of data and enable the coordination and communication among heterogeneous tasks, executing simultaneously in an HPC system, cooperating toward a common objective.
Definition of In Situ Data Management (ISDM)

The practices, capabilities, and procedures to control the organization of data and enable the coordination and communication among heterogeneous tasks, executing simultaneously in an HPC system, cooperating toward a common objective.
In Situ Data Management and Workflows

In Situ Workflows

Applications of ISDM

Workflow Management Systems (WMS)

Software frameworks for workflows

In Situ Data Management (ISDM)

Fundamental capabilities (building blocks) of in situ processing
Why In Situ?

• ISDM can make critical contributions to managing and reducing large data volumes from computations and experiments.

  Successful ISDM can minimize data movement, save storage space, and boost resource efficiency—often while simultaneously increasing scientific precision.

• The in situ methodology enables scientific discovery from a broad range of data sources, over a wide scale of computing platforms.

  Successful ISDM will benefit real-time decision making, design optimization, and data-driven scientific discovery.
Overview

- Diversity of Science Applications
- Complexity of Data Analysis Techniques
- Range of Analysis Tools
- New Science Opportunities
  - Smart Simulations
  - Ensemble Analysis, UQ
  - ML / AI
  - Surrogate Models
- New Challenges
  - Temporal Analysis
  - Human Interaction
  - Software Complexity

Application Drivers
- Traditional Uses
- I/O Staging
- Scientific Visualization
- Architecture Drivers

I/O Bottlenecks
- Hardware Heterogeneity
- Multiple Software Stacks

Time
- Past Uses
- Current Research Drivers
- Future Opportunities and Challenges
In Situ Yesterday

Simulation → Visualization

[Zajac, 1964]
In Situ Yesterday

Simulation → Visualization

[Parker et al., 1995]
In Situ Today

BES workflow of dynamic ensemble of simulations and in situ detection of stochastic events

[Yildiz et al., 2019]
Neutrino event generation and parameter optimization for DUNE (2026).
Related

Research Infrastructures, including e-Infrastructures

State-of-the-art research infrastructures become increasingly complex and costly, often requiring integration of different equipment, data sources, as well as extensive transnational collaboration.

Dear Colleague Letter: Request for Information on Data-Focused Cyberinfrastructure Needed to Support Future Data-Intensive Science and Engineering Research

October 22, 2019

Research Paper

Big data and extreme-scale computing: Pathways to Convergence-Toward a shaping strategy for a future software and data ecosystem for scientific inquiry

Priority Research Directions

“Oh the things you can find, if you don’t stay behind!”
– Dr. Seuss
Priority Research Directions

Pervasive, controllable, composable, and transparent ISDM, co-designed with the software stack and with fundamentally new algorithms.

- Pervasive
- Transparent
- Co-designed
- In Situ Algorithms
- Composable
- Controllable

Increase confidence in reproducible science, repeatable performance, and feature discovery through provenance.

Develop interoperable ISDM components for agile and sustainable programming.

Understand the design space of autonomous decision-making and control of in situ workflows.

Apply ISDM and in situ workflows at a variety of platforms and scales.

Coordinate ISDM development with underlying system software.

Redesign analysis algorithms for the in situ paradigm.

Increase confidence in reproducible science, repeatable performance, and feature discovery through provenance.

Develop interoperable ISDM components for agile and sustainable programming.

Understand the design space of autonomous decision-making and control of in situ workflows.

Increase confidence in reproducible science, repeatable performance, and feature discovery through provenance.

Develop interoperable ISDM components for agile and sustainable programming.

Understand the design space of autonomous decision-making and control of in situ workflows.
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?
Co-designed ISDM

Coordinate the development of ISDM with the underlying system software so that it is part of the software stack.

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 high-performance computing and other software stacks.
How should in situ algorithms be designed to make the 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?
In Situ Algorithms

Redesign data analysis algorithms for the in situ paradigm.

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.

In situ topological feature detection in turbulent combustion simulations used to segment and track localized intermittent ignition and extinction features. (images courtesy of J.H. Chen).
What metrics best describe the ISDM design space? How can that space be defined, codified, and evaluated to support design decision-making and control?
Controllable ISDM

Understand the design space of autonomous decision-making and control of in situ workflows.

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.

Model of how information flows for experimental computing, illustrating how real-time data analysis is required to guide the detector system, readout system, and data handling (image courtesy of Amber Boehnlein).
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?
Composable ISDM

Develop interoperable ISDM components and capabilities for an agile and sustainable programming paradigm.

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.
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?
Transparent ISDM

Increase confidence in reproducible science, deliver repeatable performance, and discover new data features through the provenance of ISDM.

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.

Performance provenance for NWChem computational chemistry simulation (image courtesy of Huub Van Dam, Wei Xu, Cong Xie, and Wonyong Jeong).
How can ISDM methodologies help meet the needs for real-time, high-velocity data applications at the edge and other non-high-performance computing platforms? How can ISDM enable science at experimental and observational facilities?
Apply ISDM methodologies and in situ workflows at a variety of platforms and scales.

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.
Process

“Plans are of little importance, but planning is essential.”
– Winston Churchill
Before the Workshop

- **Jun. 2018**: Begin discussions
- **Jul. 2018**: Select organizing committee (OC)
- **Aug. 2018**: Begin regular biweekly calls, define topics
- **Sept. 2018**: Refine scope, brainstorm invitees
- **Oct. 2018**: Draft agenda
- **Nov. 2018**: Draft pre-workshop document
- **Dec. 2018**: Finalize agenda, pre-workshop document, plenary speakers, invited participants, workshop working documents, contingency plans
- **January 2019**: Workshop
After the Workshop

- Workshop
- February
- OC writing brochure
- April
- PRDs published in brochure
- June
- Final report published
- ASCAC
- September
- Journal article submitted
- WORKS keynote
- November
At the Workshop

Science Applications
Interface to applications and science workflows

- **Data Models: Connection and Communication**
  Structure, semantics, and movement of in situ data

- **Computational Platforms and Environments**
  Interface to hardware and system software stacks and future platforms

- **Analysis Algorithms**
  Portable, high-performance algorithms that can be used in situ and elsewhere

- **Provenance and Reproducibility**
  Information for diagnostics, performance studies, and scientific reproducibility

- **Programming and Execution Models**
  Programming and executing disparate constituent tasks in an ISDM framework

- **Software Architecture for Usability and Sustainability**
  Software that can be built, deployed, sustained, and used to support DOE science
From Workshop Topics to PRDs
<table>
<thead>
<tr>
<th>Topic</th>
<th>Subtopic</th>
<th>PRD→</th>
<th>Pervasive</th>
<th>Codesigned</th>
<th>Algorithms</th>
<th>Controllable</th>
<th>Composable</th>
<th>Transparent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Models</td>
<td>Multimodal Science</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Expressing Intent</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Computational Platforms</td>
<td>System Software</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Hardware Co-design</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Analysis Algorithms</td>
<td>Reduced Representation</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Decision Making</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>New Outputs</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Provenance / Reproducibility</td>
<td>Capture</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Processing</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Reproducible Science</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Programming / Execution Models</td>
<td>Elasticity</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Optimizing Execution</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Composability</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Streaming</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Software Architecture</td>
<td>Use Cases</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Usability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Interoperability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Confidence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Science Facilities</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
### Key Challenges and Opportunities

Please describe the underlying science challenges and opportunities that motivate this PRD.

### State of the art

Please answer the following questions:
- Who else is doing this?
- What are the technology and research gaps?

### New Research Direction

Please answer the following questions:
- What will you do to address the challenge?
- What research questions will you ask / answer?
- What are the potential risks?
- What would success look like?
- What assumptions about users, hardware, or other parts of the software stack motivate this as a priority / are required for success?

### Potential Scientific Impact

Please answer the following questions:
- What new scientific capabilities will follow?
- What new methods and techniques will be developed?
Data Capture

Key Challenges and Opportunities
Please describe the underlying science challenges and opportunities that motivate this PRD

State of the art
Please answer the following questions:
• Who else is doing this?
• What are the technology and research gaps?

New Research Direction
Please answer the following questions:
• What will you do to address the challenge?
• What research questions will you ask / answer?
• What are the potential risks?
• What would success look like?
• What assumptions about users, hardware, or other parts of the software stack motivate this as a priority / are required for success?

Potential Scientific Impact
Please answer the following questions:
• What new scientific capabilities will follow?
• What new methods and techniques will be developed?
Assumptions and Dependencies Matter

• Track the consequence of assumptions back to priorities if/when assumptions change.
• Follow relationships between parts of the research portfolio.
• Promotes a software stack view of the portfolio.
• Components of the portfolio can work together to achieve capabilities.
Day 1: Parallel Breakout Sessions

- Breakout session 1A
- Breakout session 2A
- Breakout session 3B

Research Area 1A-I
Research Area 2A-I
Research Area 3B-I

Drafting PRDs

Evening 1 - Morning 2: Draft PRD Synthesis

Organizing Committee

Day 2: PRD Report-Back and Discussion

- Draft PRD 1
- Draft PRD 2
- Draft PRD 3
- Draft PRD 4
- Draft PRD 5
- Draft PRD 6

ISDM Workshop Report
Take-Aways

Decouple discussion topics and PRDs.

- Discussion topics are textbook chapters, college courses
- PRDs are desired capabilities
- They don’t map 1:1
- You know the discussion topics in advance
- PRDs come out of the workshop
Take-Aways

Draft PRDs at the workshop.

• Review with participants and get their feedback
• Rest of the report writing follows
Thank You

Organizing Committee
Debbie Bard, Janine Bennett, Wes Bethel, Ron Oldfield, Line Pouchard, Christine Sweeney, and Matthew Wolf

Program Manager
Laura Biven

Workshop Participants
Resources

Brochure (4 pages)
Full report (100 pages)

DOI: 10.2172/1493245

https://science.osti.gov/ascr/Community-Resources/Program-Documents