Satisfying the data demands of DOE science

Rob Ross, Philip Carns, Matthieu Dorier, Kevin Harms, Robert Latham, Pierre Matri, and Shane Snyder
Argonne National Laboratory

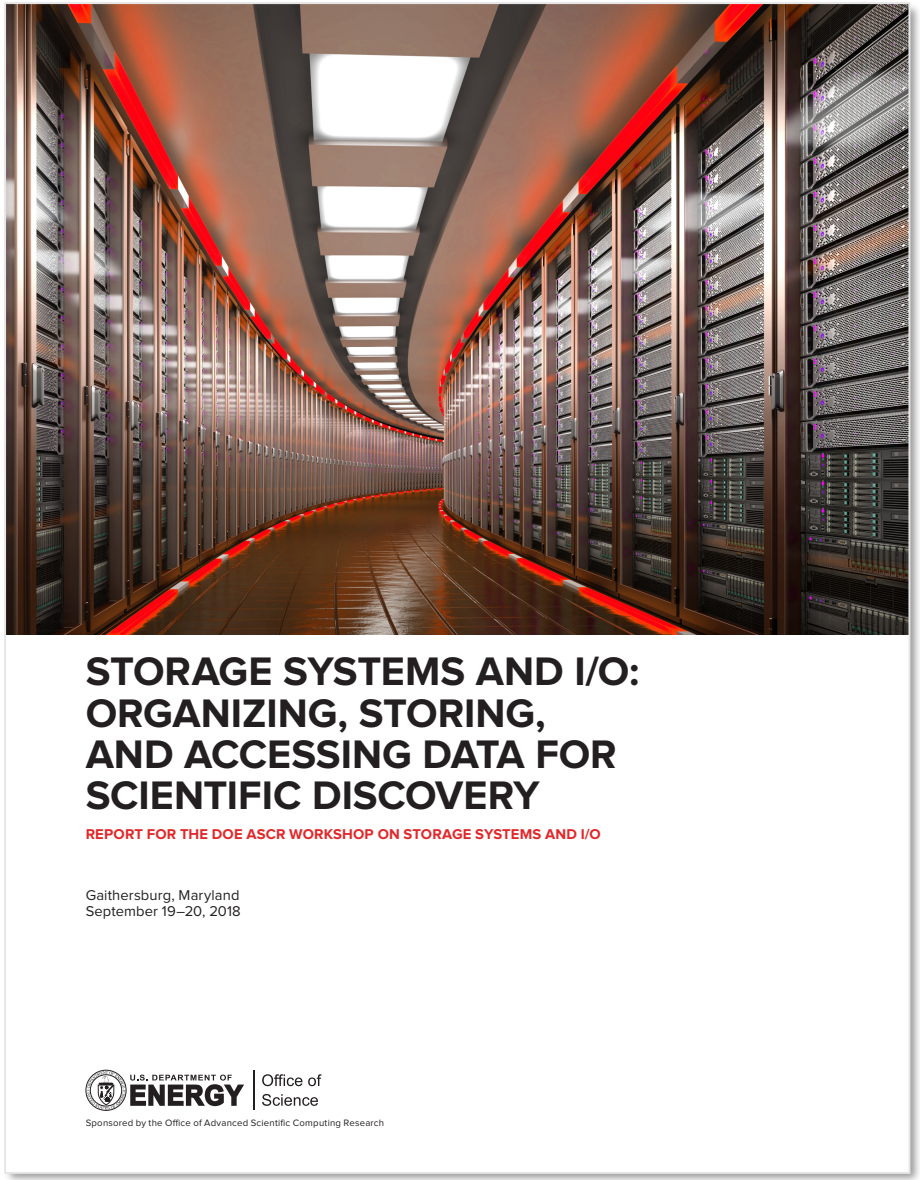
Bob Robey, Brad Settlemyer, and Galen Shipman
Los Alamos National Laboratory

George Amvrosiadis, Chuck Cranor and Greg Ganger
Carnegie Mellon University

Neelam Bagha and Jerome Soumagne
The HDF Group

2018: Application Drivers of Data Research and Development

- 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



2018: Application Drivers of Data Research and Development

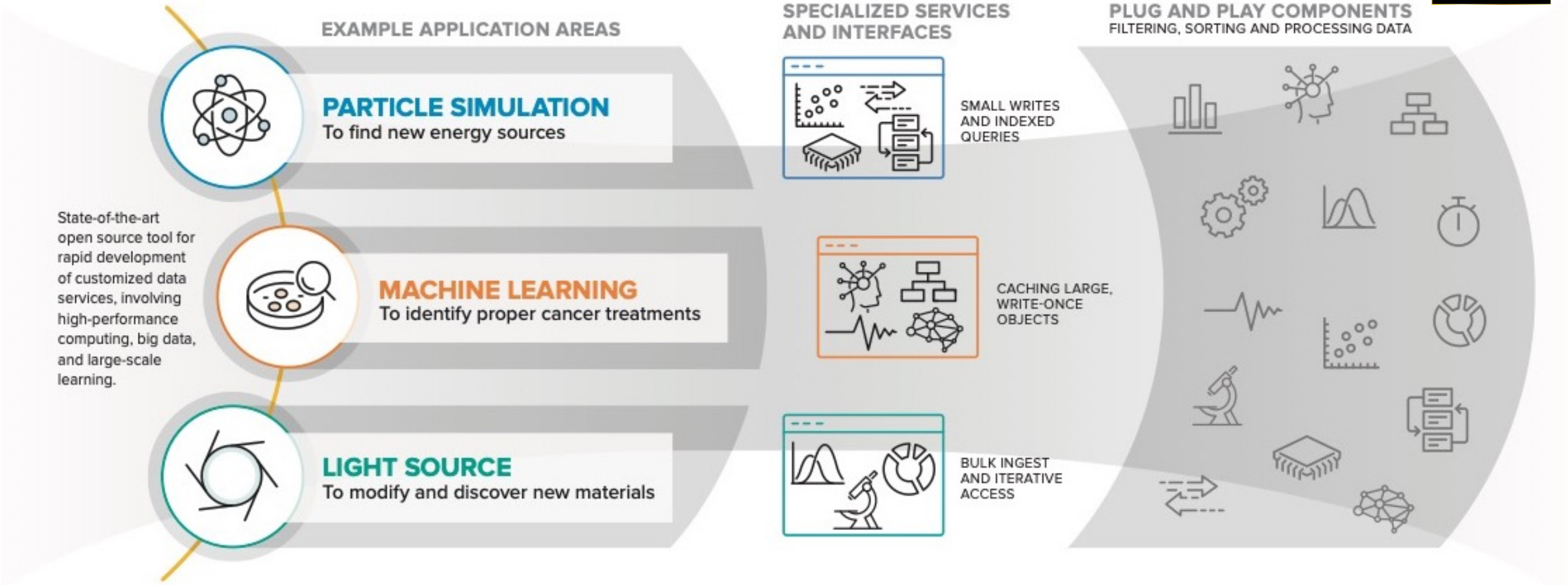
- Experimental/observational data science
  - **Streaming data** model
  - **Many small records**
- ***DOE science exhibits many different types of data (data models), ways of accessing (interfaces), and patterns of access (organizations).***
- In situ data analysis
  - **Fine grained sharing** between tasks
  - High penalty for data transformations
- Reproducibility
  - Greater reliance on provenance information

Bespoke data services:
The Mochi project

Carnegie Mellon University
HDF
Los Alamos NATIONAL LABORATORY EST. 1943
Argonne NATIONAL LABORATORY

Mochi: a Data Service Development Environment

2021 R&D 100 FINALIST

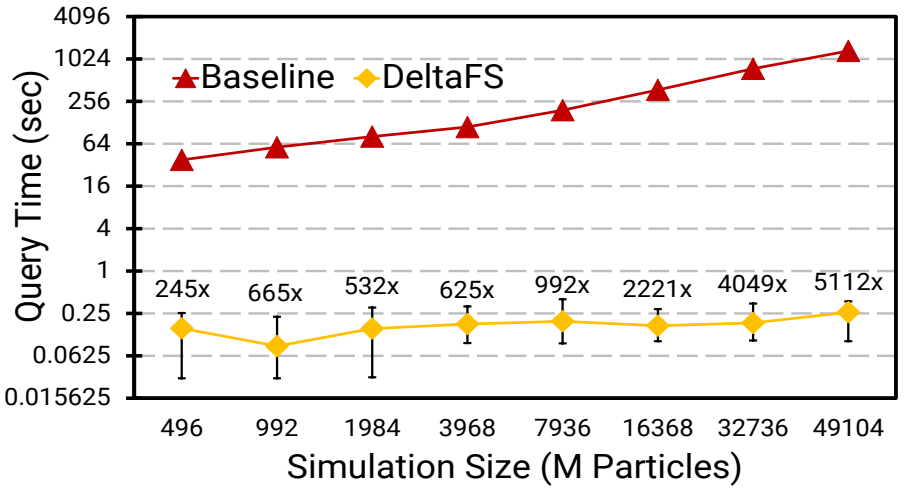


Carnegie Mellon University · HDF · Los Alamos National Laboratory EST. 1943 · Argonne National Laboratory

Components of data services

	Component	Summary
Core	**Argobots**	Argobots provides user-level thread capabilities for managing concurrency.
	Mercury	Mercury is a library implementing remote procedure calls (RPCs).
	Margo	Margo is a C library using Argobots to simplify building RPC-based services.
	Thallium	Thallium allows development of Mochi services using modern C++.
	SSG	SSG provides tools for managing groups of providers in Mochi.
Utilities	**ABT-IO**	ABT-IO enables POSIX file access with the Mochi framework.
	Bedrock	Bedrock is a bootstrapping and configuration system for Mochi components.
	ch_placement	ch-placement is a library implementing multiple hashing algorithms.
	Shuffle	Shuffle provides a scalable all-to-all data shuffling service.
Microservices	**BAKE**	Bake enables remote storage and retrieval of named blobs of data.
	Neon	Neon is a microservice for building streaming data services.
	POESIE	Poesie embeds language interpreters in Mochi services.
	REMI	REMI is a microservice that handles migrating sets of files between nodes.
	SDSKV	SDSKV enables RPC-based access to multiple key-value backends.
	Sonata	Sonata is a Mochi service for JSON document storage based on UnQLite.

A Data Service for Kinetic-Plasma Simulations (DeltaFS)



Indexing during post-processing is becoming increasingly time consuming for exascale applications. DeltaFS uses spare resources on compute nodes to perform in situ indexing. The result is 3+ orders of magnitude improvement in query speed.

- VPIC is a first-principles, kinetic-plasma simulations code
  - Simulations often model trillions of particles and require months of run time
- An important analysis is to examine the history of high-energy particles identified at end of run
- DeltaFS in situ indexing facilitates fast generation of these histories, enabling interactive query (left)
  - Leverages Mochi components and LevelDB technology from Google

2019 R&D 100 WINNER

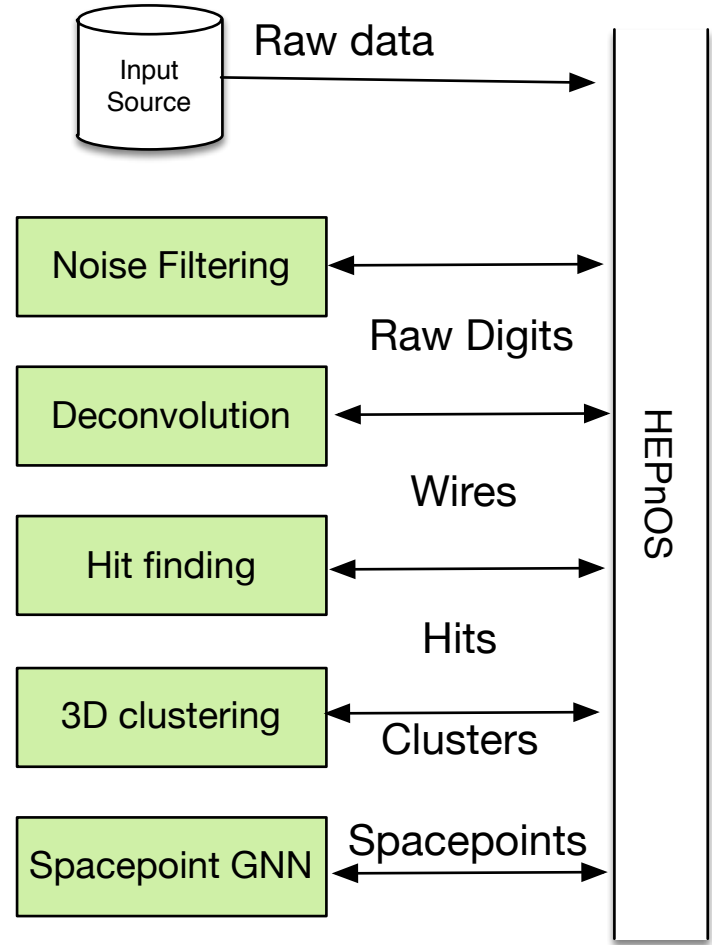
Carnegie Mellon University · HDF · Los Alamos National Laboratory EST. 1943 · Argonne National Laboratory

High-Energy Physics Event Store (HEPnOS)

When moving HEP analysis to HPC platforms, the parallel file system is a bottleneck due to many small accesses to events and products.

- HEPnOS is a temporary store for event data
- Integrates easily with analysis tools (e.g., *art*)
- Accelerates put/get of event data, scales

- Demonstrating with NuMI Off-Axis Electron Neutrino Appearance (NOvA) 4th analysis workflow
- Extending to support Imaging Cosmic And Rare Underground Signals (ICARUS) workflow (right)



HEPnOS in the ICARUS workflow.

Fermilab Carnegie Mellon University HDF Los Alamos National Laboratory Argonne National Laboratory

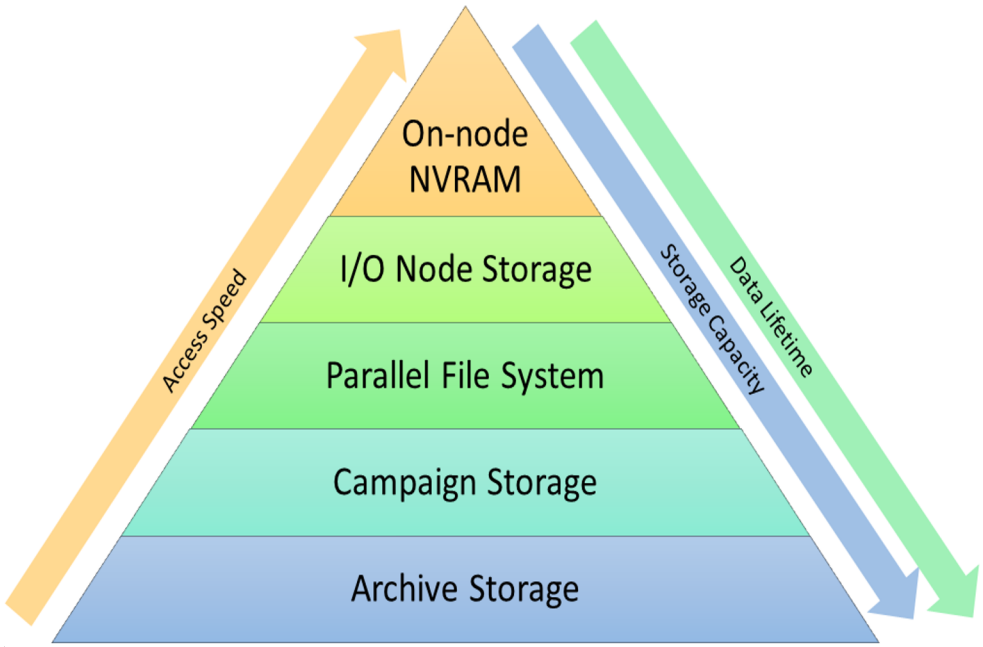
Other data services built with Mochi technologies

Service	Institution(s)	Summary
Chimbuko	Brookhaven	Workflow-level scalable **performance trace analysis** tool
DAOS	**Intel**	**Object store** providing high bandwidth, low latency, and high I/O operations per second (IOPS) storage containers to HPC applications
DataSpaces	Univ. of Utah	Programming system and data management framework for **coupled workflows**
GekkoFS	Univ. of Mainz	Temporary distributed file system for HPC applications
Hermes	IIT, THG, UIUC	User-space platform for **distributing data structures**
HXHIM	Los Alamos	**Hexadimensional hashing indexing** middleware
Proactive Data Containers	Berkeley	**Object-centric data management** system to take advantage of deep memory and storage hierarchy
Seer	Los Alamos	Lightweight **in situ** wrapper library adding in situ capabilities to simulations
Unify	LLNL and ORNL	Suite of specialized**, flexible file systems** that can be included in a user's job
	Kitware	Platform for ubiquitous access to **visualization results during runtime**
Dist. Systems Coursework	Tsukuba	Creating distributed systems using RPC and RDMA as part of an **Information Systems course**

Looking Ahead:
Technologies, AI, and Data Management

Carnegie Mellon University
HDF
Los Alamos NATIONAL LABORATORY EST. 1943
Argonne NATIONAL LABORATORY

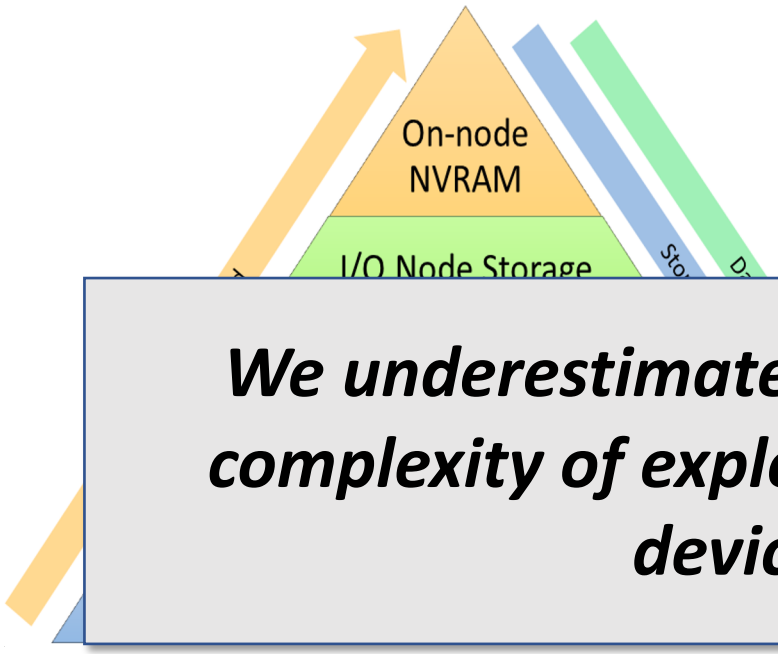
2018: Technology Drivers of Data Research and Development



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.

- 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

2018: Technology Drivers of Data Research and Development



- New memory and storage technologies
  - **Blurring lines between storage and memory**
  - New access methods

We underestimated the potential for nonvolatile memory and the complexity of exploiting it, and we were too skeptical that "smart" devices would appear in the near term.

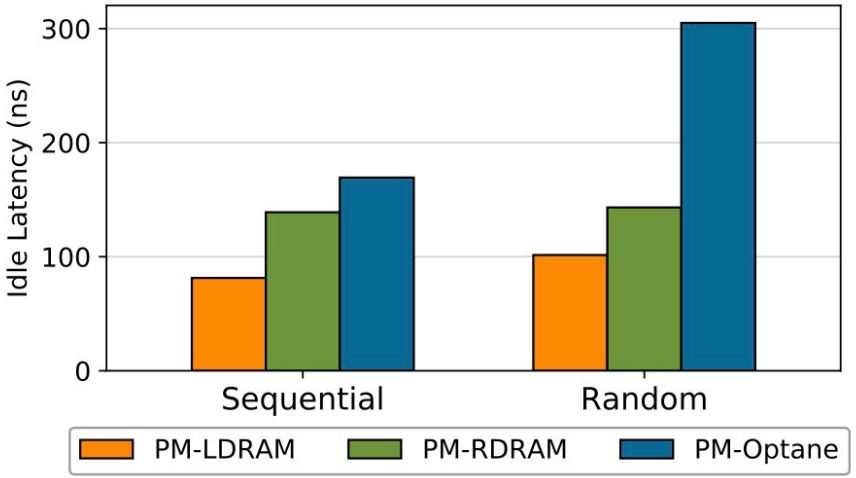
New memory and storage technologies provide opportunities to retain more data than ever before, to directly and efficiently access individual records regardless of location in the system, and to lower costs by employing the most economically viable technologies for specific tasks.

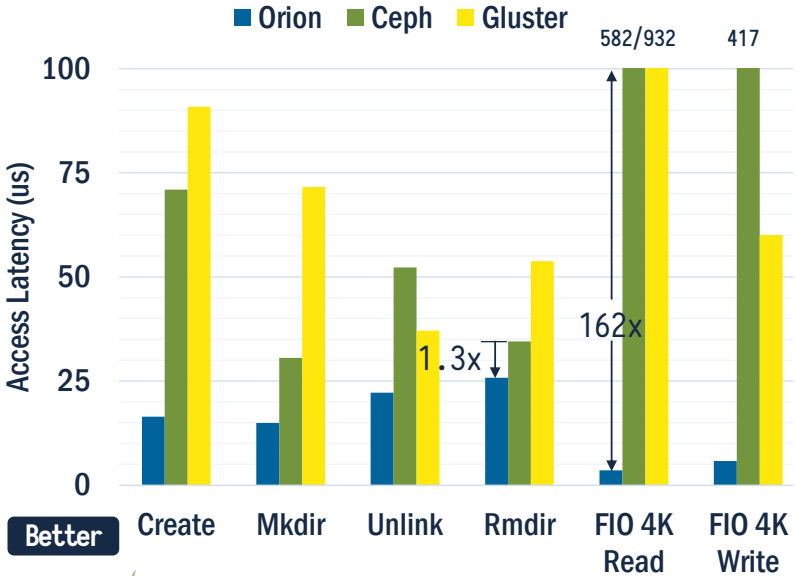
- Deeper storage hierarchy than in the past
  - Positioning and locating data more difficult
  - Widely varying performance characteristics
- Interconnects with new characteristics
  - Emerging quality of service features

IOPS and why they matter

IOPS. *noun.* **input/output operations per second**; a measure of performance for storage, memory, and networking devices.

- Data analysis and learning workloads have shifted storage access towards more reads and many small accesses
- High speed networks and nonvolatile memory provide the basis for solutions
- Research needed to understand how to streamline the fast paths for our services





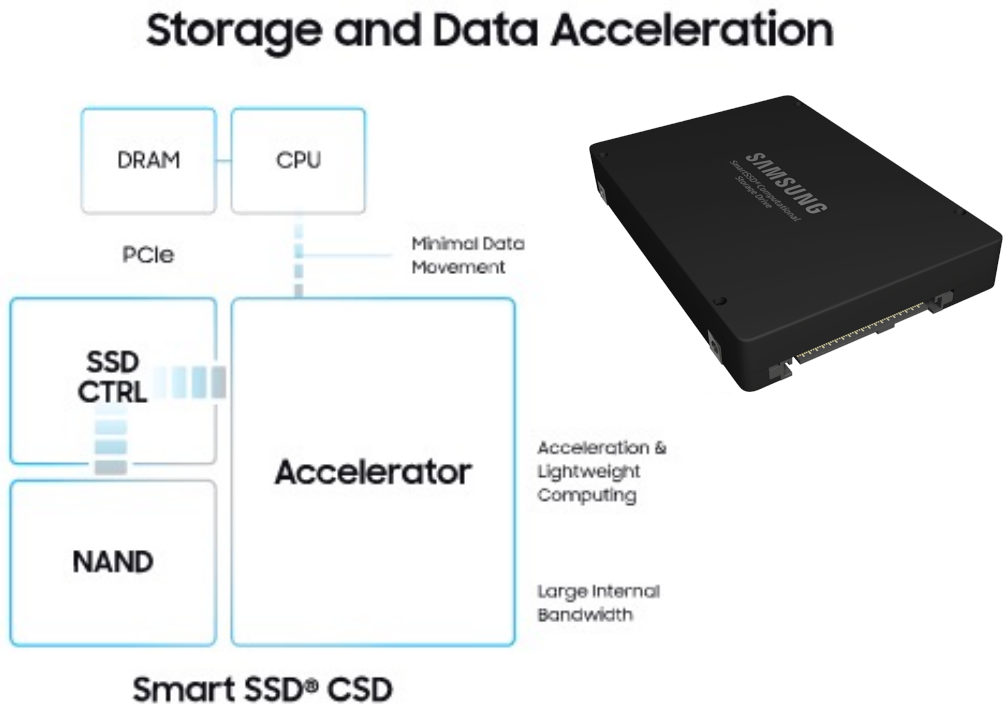
Hardware capabilities of new devices are not being fully exploited, even by leading research efforts.

Sources: J. Izraelevitz et al., "Basic Performance Measurements of the Intel Optane DC Persistent Memory Module," arxiv, August 2019 (top) and J. Yang et al. "Orion: A Distributed File System for Non-Volatile Main Memories and RDMA-Capable Networks," FAST, February 2019 (bottom).

Carnegie Mellon University · HDF · Los Alamos National Laboratory · Argonne National Laboratory

Smart devices

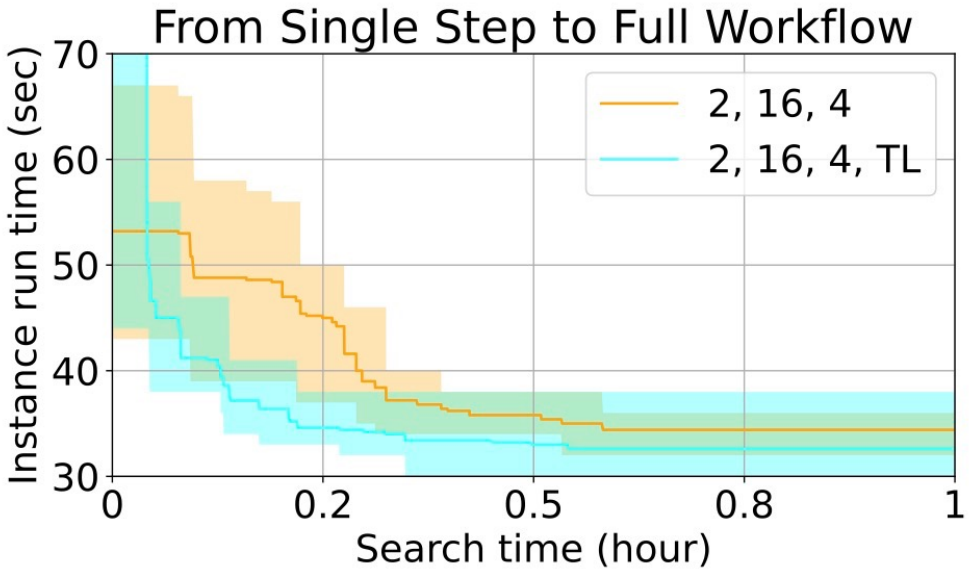
- Computational capabilities integrated in network and storage devices
- HPC has found limited uses in the past
  e.g., accelerating collective communication
- Hardware trends mean these are much closer to CPU capabilities than before!
- Potentially aid in removing software from the fast path?
  Offload service capabilities to device rather than application tasks?



Samsung SmartSSD computational storage drive (CSD) incorporates Xilinx compute capabilities into the device, enabling computing in storage and other forms of customization of device behavior and interfaces.

Learning algorithms, AI, and data research

- **Accelerating AI on HPC:** Faster access to training data, faster checkpoint/ restart, performance tuning workloads
- **Data specialization and augmentation:** New data formats and services that match AI workflow needs, capture of information enabling FAIR models
- **AI for services:** Improving initial configurations (right), adapting to changing workloads and environments, detecting anomalies



Searching for good HEPnOS configurations. Transfer learning allows us to leverage knowledge from prior experiments (one workflow step) to accelerate search for good configurations for the full workflow.

Carnegie Mellon University · HDF · Los Alamos National Laboratory EST. 1943 · Argonne National Laboratory

Bridging the data “worlds”

- Massive volumes
- High throughput
- Low latency
- Parallelism
- Fault tolerance
- Programmability

Storage and I/O World

- Findability
- Accessibility
- Interoperability
- Reusability
- Ontologies
- Data life cycle

Data Management World

https://clipartix.com/world-clip-art-image-53181/

Carnegie Mellon University HDF Los Alamos NATIONAL LABORATORY EST. 1943 Argonne NATIONAL LABORATORY

Bridging the data “worlds”

DOE science requirements span these worlds! Rich opportunities lie in combining these concepts and technologies:

- ***Annotations capturing workflow details for reproducibility***
- ***“Smart stores” connecting relevant datasets at exabyte scale***
- ***“Collaborative stores” sharing information across institutional boundaries***

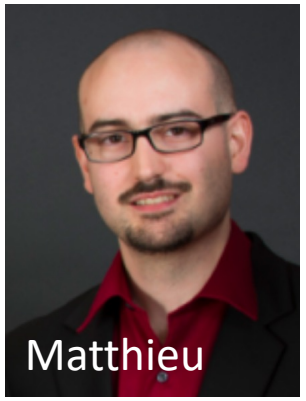
Storage and I/O World

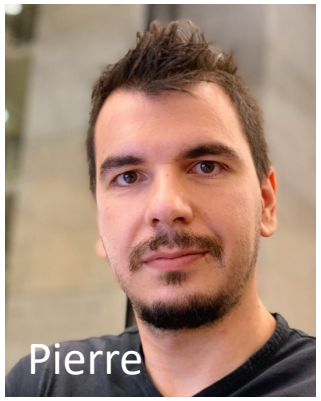
Data Management World

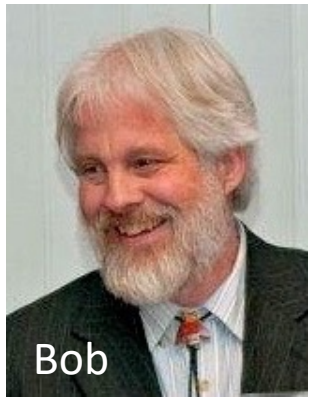
https://clipartix.com/world-clip-art-image-53181/

Carnegie Mellon University HDF Los Alamos NATIONAL LABORATORY EST. 1943 Argonne NATIONAL LABORATORY

Data and storage research continues to pay big dividends for DOE science and facilities, with plenty of interesting challenges ahead!







Carnegie Mellon University

HDF

Los Alamos NATIONAL LABORATORY EST. 1943

Argonne NATIONAL LABORATORY