DataLib: Data Libraries and Services

PI:	Rob Ross	(ANL)
Co-PIs:	Galen Shipman	(LANL)
	Wei-keng Liao	(Northwestern)
	Jerome Soumagne	(The HDF Group)

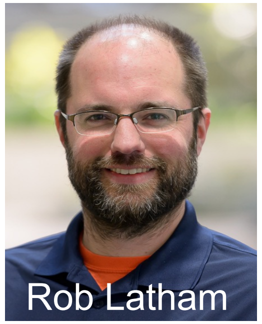
March 30, 2021

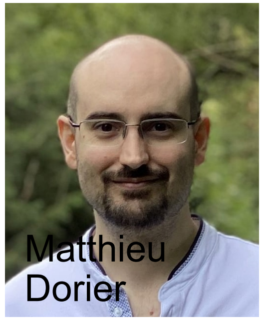
EXASCALE COMPUTING PROJECT

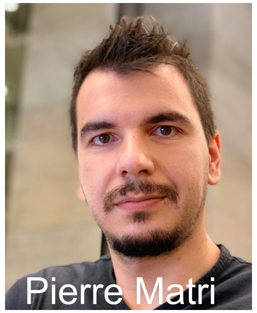
National Nuclear Security Administration

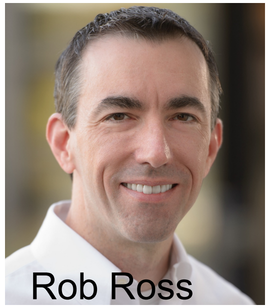
U.S. DEPARTMENT OF ENERGY | Office of Science

The DataLib Team

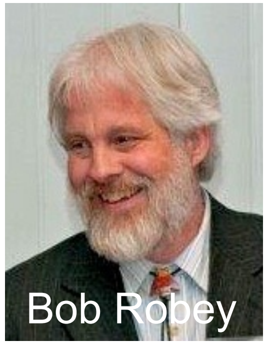




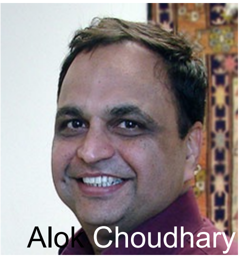




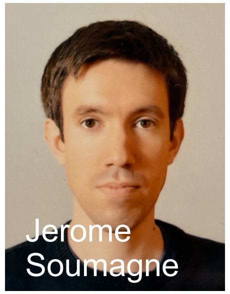
Argonne

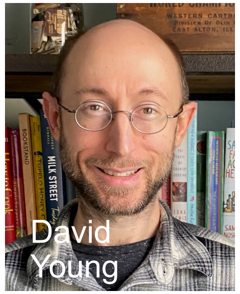


Los Alamos



Northwestern





The HDF Group

DataLib Strategy

User-level storage and I/O for ECP codes on upcoming DOE platforms

Members responsible for some of the most successful storage and I/O software in the DOE complex.

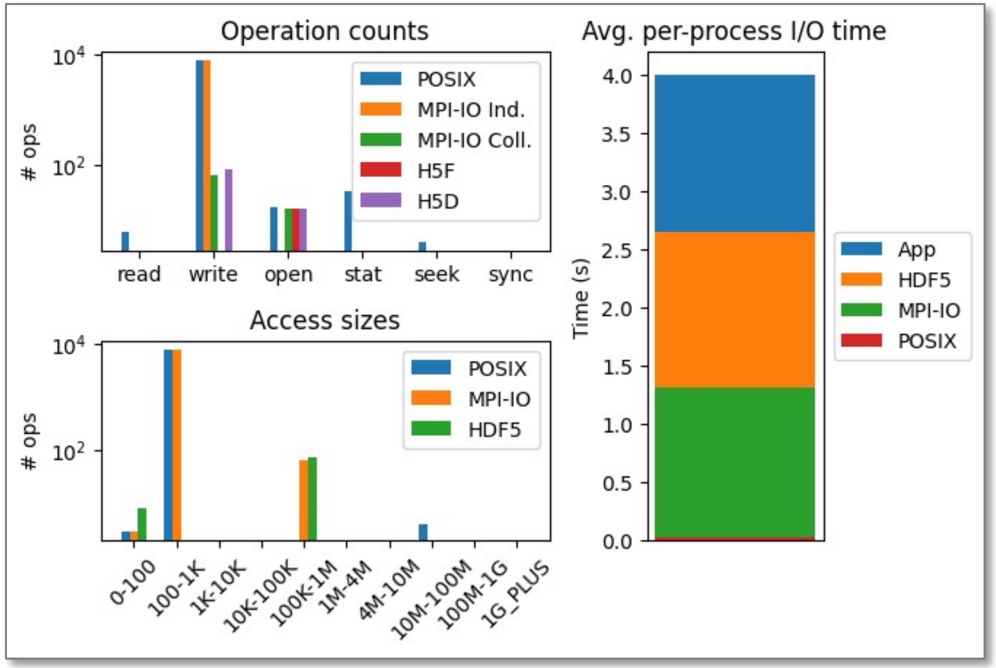
Our software:

- **Darshan.** Lightweight I/O characterization for HPC codes
- **ROMIO and Parallel netCDF.** Standards-based I/O for HPC
- **Mochi.** Customized data services for DOE science
- **Datalib HDF5 VOL.** Accelerated I/O for HDF5 users

This talk will briefly introduce each of these tools.

Darshan
Lightweight I/O characterization for HPC codes

- **Darshan** is a tool for observing application I/O patterns on production HPC platforms, typically installed by facility operators and enabled by default.
- Who uses Darshan?
  - Facilities looking to gain greater insight into their users' I/O behavior
  - Application teams looking to understand I/O bottlenecks
  - Performance engineers helping teams maximize I/O productivity
- What's new?
  - HDF5, Parallel netCDF, and DAOS modules provide greater detail on these interfaces
  - New python analysis tools help understand results



Darshan data provides a view of I/O behavior at multiple levels. In this MACSio example we can see that significant time is spent in HDF5 and MPI-IO. If performance tuning were undertaken, those two layers would be the initial focus.

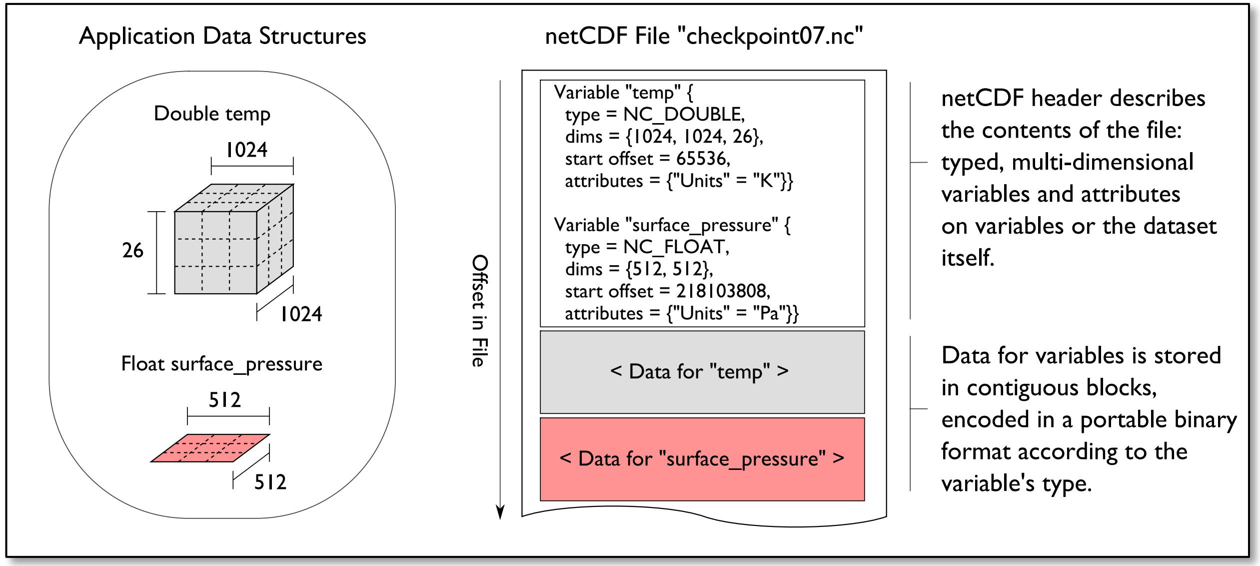
EXASCALE COMPUTING PROJECT

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

ROMIO and Parallel netCDF

Standards-based I/O for HPC

- **ROMIO** is an implementation of the I/O part of the MPI standard, included in MPICH and many vendor-supplied MPI implementations.
- **Parallel netCDF** is a portable API and format for storing and sharing scientific data, modeled after the netCDF-3 interface.
- Who uses ROMIO and PnetCDF?
  - End users and I/O library writers employ ROMIO as a portable I/O interface for "low level" file system access.
  - End users employ PnetCDF as an efficient and descriptive scientific data format.
- What's new?
  - Numerous performance optimizations for current and future platforms.



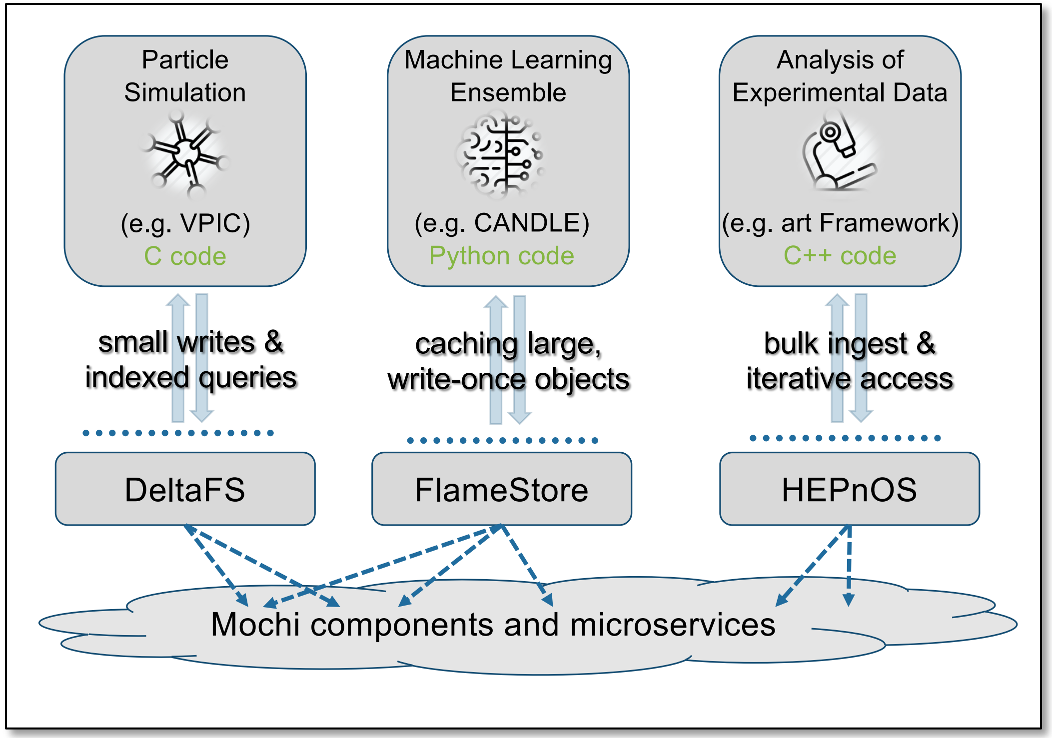
Libraries such as Parallel netCDF, HDF5, netCDF-4, and ADIOS provide mechanisms to not only store data but to describe the structure of that data and to capture significant provenance.

While these libraries can sometimes exhibit lower performance than more "bare metal" use of file systems, we strongly encourage science teams to consider these libraries as a way to improve their productivity and the portability of their data.

https://parallel-netcdf.github.io/
https://www.mcs.anl.gov/projects/romio/

Mochi
Customized data services for DOE science

- **Mochi** provides a toolkit for building high-performance data services for use on HPC platforms, and ECP computer scientists are using Mochi to build services for ECP application teams.
- Who uses Mochi?
  - Computer scientists use Mochi to develop customized data services.
  - End users benefit from the specialization of these services in terms of ease of use and performance.
- What's new?
  - The Bedrock component enables easier configuration of multi-component deployments on single nodes.
  - SSG improvements have made group membership more robust.



Mochi has been used to develop a number of services, including ones to store and index particle data, to manage learning data, and to provide fast access to high-energy physics detector data during analysis.

Within ECP, Mochi is also helping enable Unify, Chimbuko, DataSpaces, and Proactive Data Containers.

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

DataLib HDF5 Plug-in
Accelerated I/O for HDF5 users

- **HDF5** is a popular choice for storing and retrieving scientific data. Our plug-in (called a VOL) will accelerate I/O for many codes.
- Who could use our HDF5 VOL?
  - When complete, any HDF5 user could switch to our VOL with few or no code changes.
  - Accelerated HDF5 will open up HDF5 use to teams who have found performance inadequate in the past.
- What's new?
  - Initial implementation being performance tuned using sample ECP use cases.
  - Performance tuning has already resulted in rates competitive with Parallel netCDF!

E3SM F case high-resolution (ne120) case study performed on Cori @ NERSC, using Lustre stripe count = 64, stripe size = 8 MiB, running 1024 MPI processes, 32 Haswell nodes

	PnetCDF	First Prototype	Latest version
File size (GiB)	14.08	91.09	22.32
Effective bandwidth (MiB/s)	745.54	62.39	1074.11
Initialization time (sec)	0.07	0.1	0.21
Time posting write requests (sec)	0.37	10.05	1.74
Time flushing the data (sec)	18.9	5.65	5.63
Time flushing the metadata (sec)	0.01	212.63	5.7
Metadata overhead (MiB)	0	76.64	8.24
End-to-end time (sec)	**19.35**	**231.23**	**13.43**