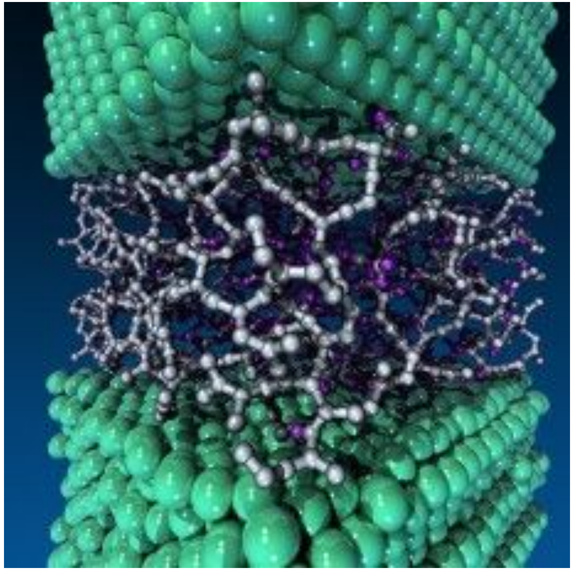
Toward Scalable and Asynchronous Object-centric Data Management for HPC

Houjun Tang, Suren Byna, Francois Tessier, Teng Wang, Bin Dong, Jingqing Mu, Quincey Koziol, Jerome Soumagne, Venkatram Vishwanath, Jialin Liu, Richard Warren

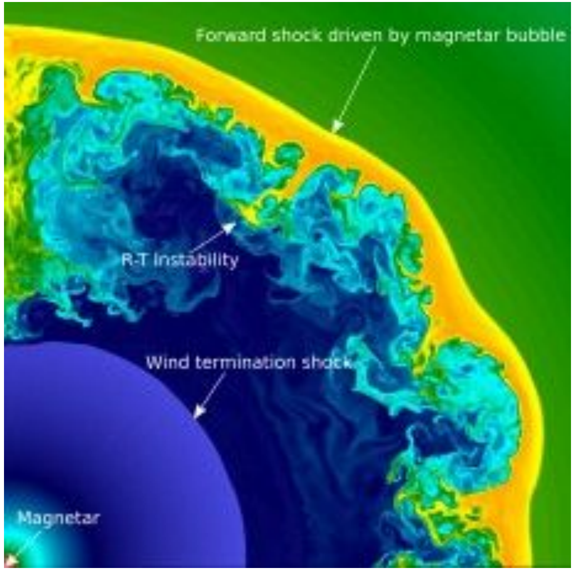
Berkeley Lab, Argonne National Lab, The HDF Group

https://sdm.lbl.gov/pdc

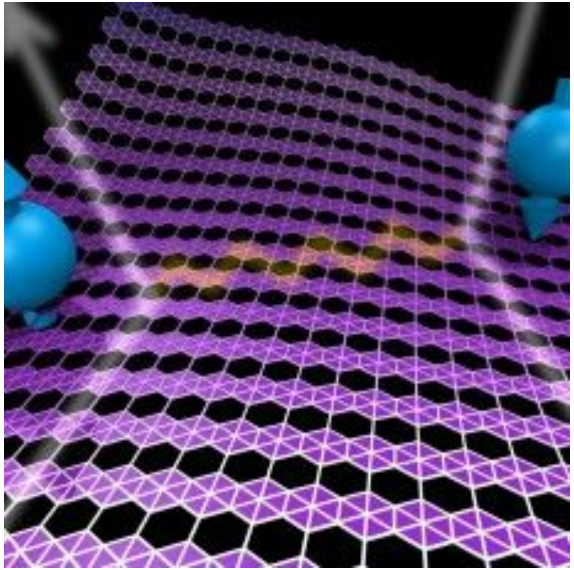
Data-driven Science



Molecular Dynamics Simulations



Superluminous Supernovae



Superconducting

Images from http://www.nersc.gov/news-publications/nersc-news

Storage Systems and I/O: Current status

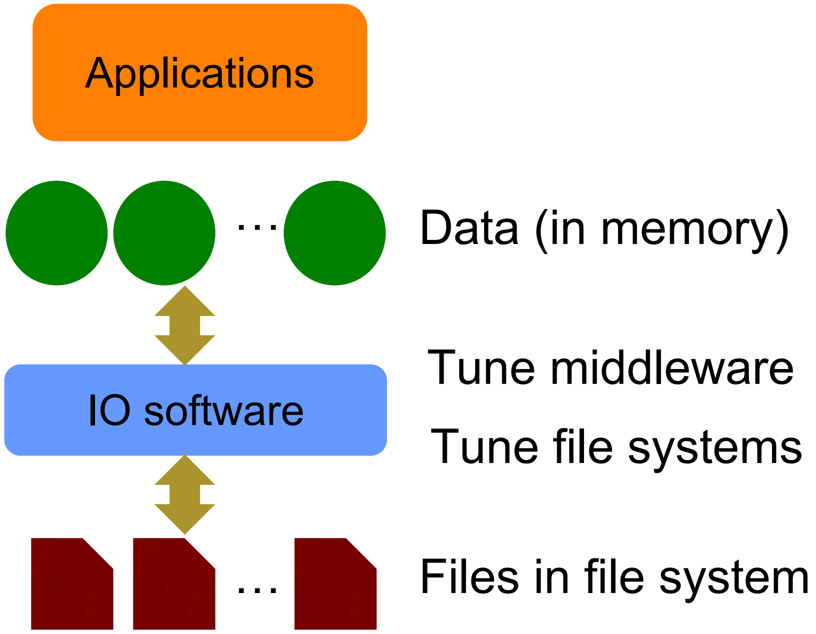
Hardware

- Memory
- Node-local storage
- Shared burst buffer
- Disk-based storage
- Campaign storage
- Archival storage (HPSS tape)

Software

- High-level lib (HDF5, etc.)
- IO middleware (POSIX, MPI-IO)
- IO forwarding
- Parallel file systems

Usage

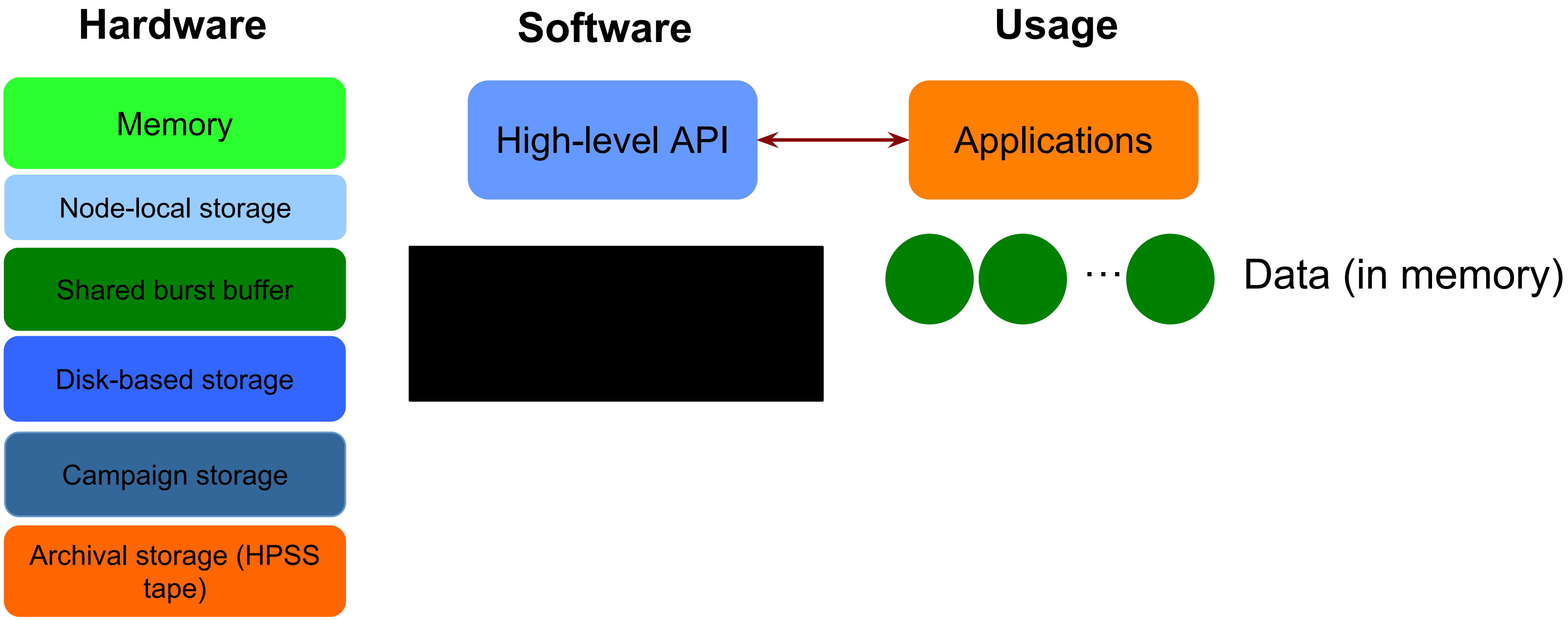


HPC I/O

- **Challenges**:
  - POSIX-IO semantics hinder **scalability** and **performance** of file systems and IO software.
  - **Multi-level hierarchy** complicates data movement, especially if user has to be involved.

- **Requirements**:
  - Simple interfaces and superior performance.
  - Autonomous data management.
  - Information capture and management.

Storage Systems and I/O: Next Generation



Storage Systems and I/O: Next Generation

- Autonomous, proactive data management system beyond POSIX restrictions.

- Transparent data object placement and organization across storage layers with tunable consistency.

- Object-centric storage with rich metadata, accessible through queries.

What is an object?

- Chunks of a file
- Files (images, videos, etc.)
- Array
- Key-value pairs
- File + Metadata

Current parallel file systems

Cloud services (S3, etc.)

HDF5, DAOS, etc.

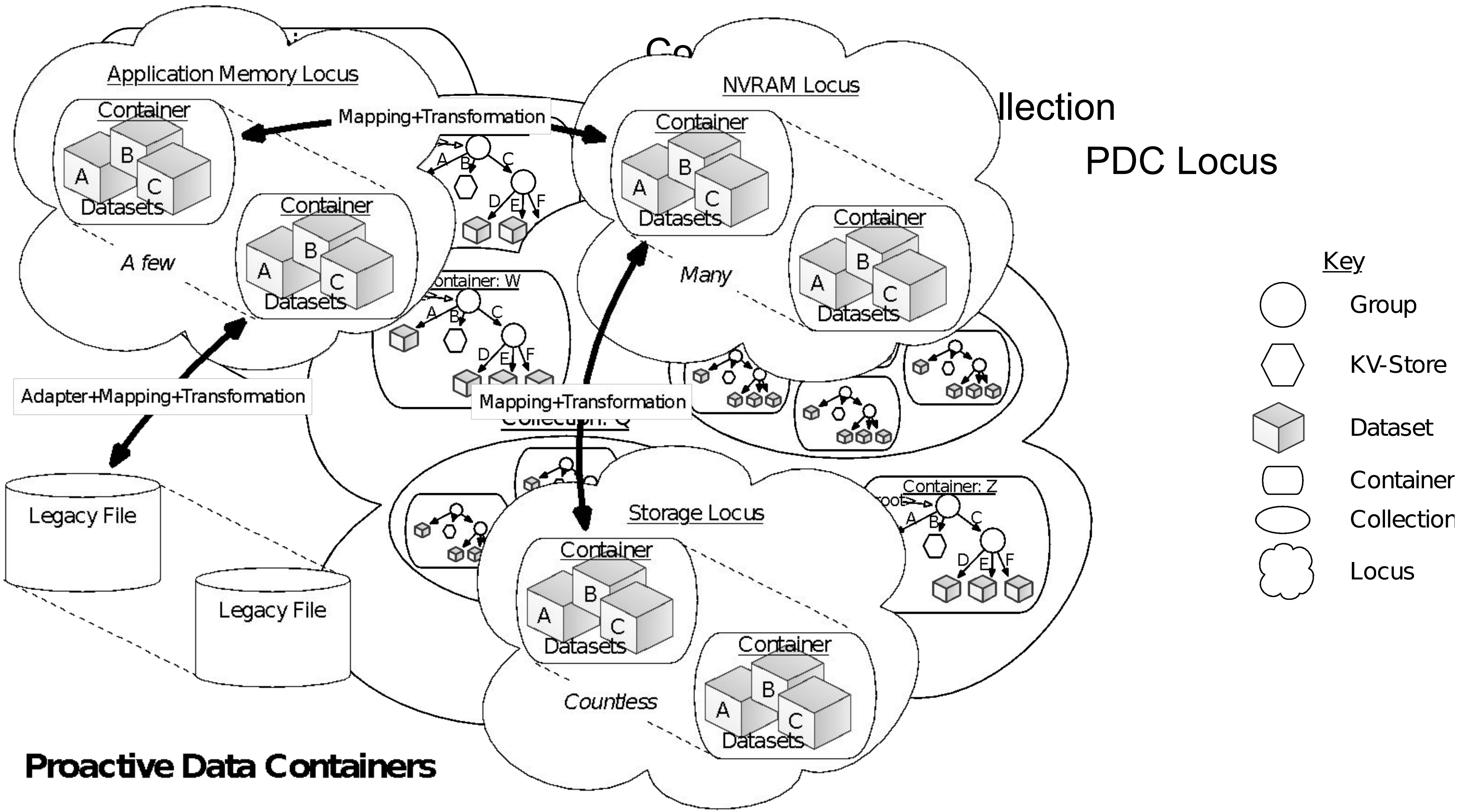
OpenStack Swift, MarFS, Ceph, etc.

PDC Interpretation of Objects

Data + Metadata + Provenance + Analysis operations + Information (data products)

↓

Proactive Data Containers (PDC)

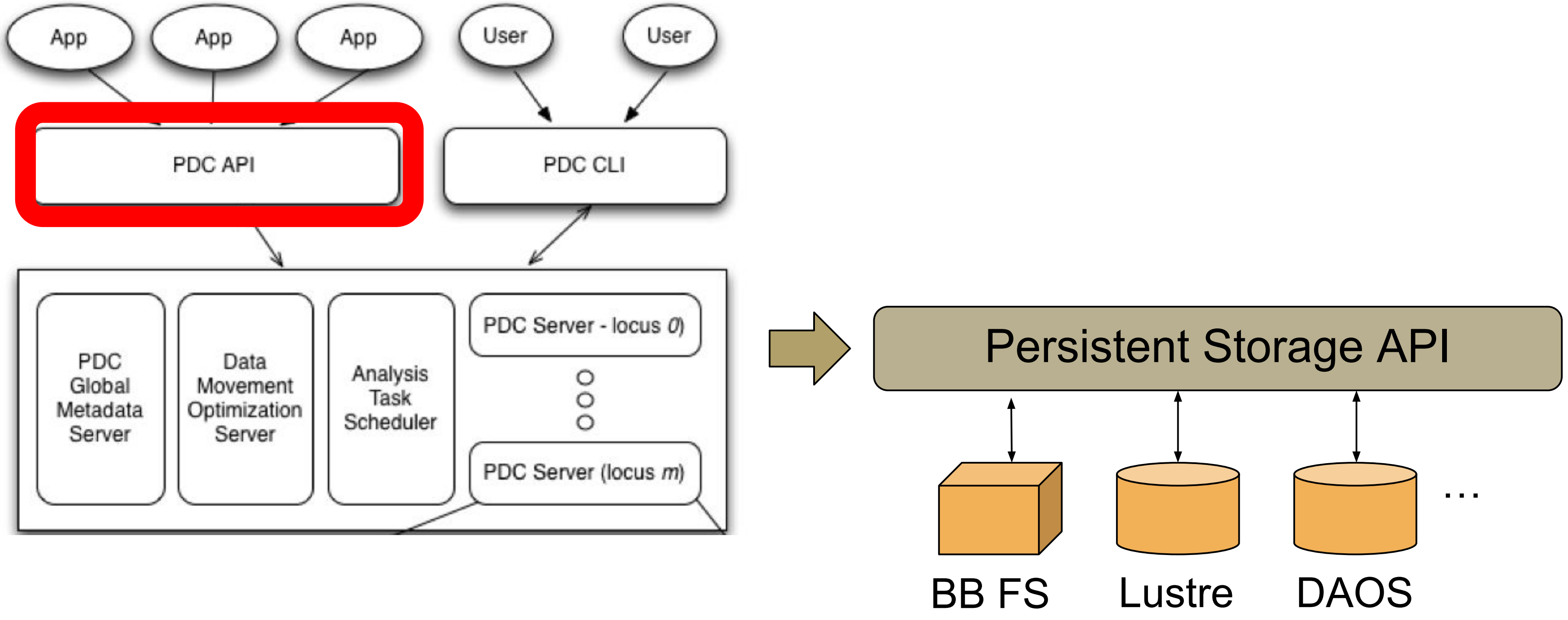


Application Memory Locus
Container
Datasets
A few
Mapping+Transformation
NVRAM Locus
Many
llection
PDC Locus
Container: W
Adapter+Mapping+Transformation
Legacy File
Legacy File
Collection: Q
Storage Locus
Countless
Container: Z
Key
Group
KV-Store
Dataset
Container
Collection
Locus
Proactive Data Containers

PDC System - High-level Architecture

- **Interface**
  - Programming and client-level interfaces
- **Services**
  - Metadata management
  - Autonomous data movement
  - Analysis and transformation task scheduler
- **PDC locus services**
  - Object mapping
  - Local metadata management
  - Locus task execution

PDC System - High-level Architecture



Object-centric API

- **Container and Object management**
  - Create and delete
- **Metadata management**
  - Set / get properties
    - Object name, dimensions, data type,
    - Analysis functions, transformations, relationships, etc.
- **I/O**
  - Put (Write)
  - Get (Read)
- **Query**
  - Metadata query
  - Data query

Metadata Management

Requirements - Efficient Metadata Management

BERKELEY LAB
Lawrence Berkeley National Laboratory

- **Scalable**
  - Effectively management of a large number of objects.
- **Extensible**
  - Attach more information anytime without a limit
- **Queryable**
  - Find interested objects by specifying a few attributes (exact or partial).

Metadata Object

A collection of ***tags*** (key-value pairs)

Pre-defined Tag	User-defined Tag

• Object ID
• DataObjLocation
• SystemInfo
• ID Attributes
  - Name - Owership
  - AppName - TimeStep
	• (UserTag1, Value1)
	• (UserTag2, Value2)
	• (UserTag3, Value3)
	• ...
	• ...

Capabilities

- Create, update, search, and delete metadata objects.
- All tags are searchable.
- Maintain extended attributes and object relationships.

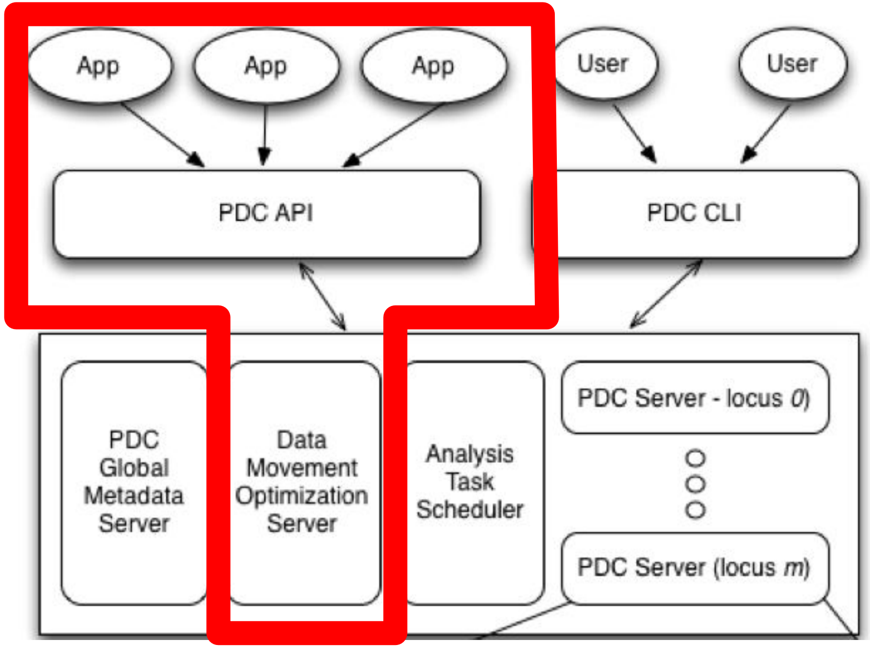
Data Movement Management

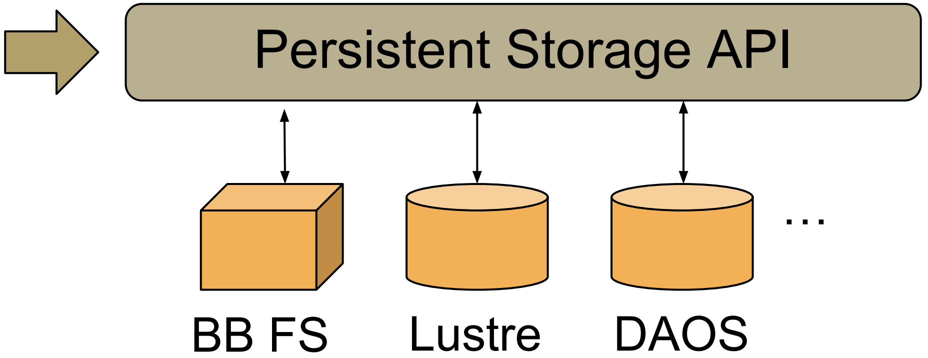
Requirement - Efficient Data Management

BERKELEY LAB
Lawrence Berkeley National Laboratory

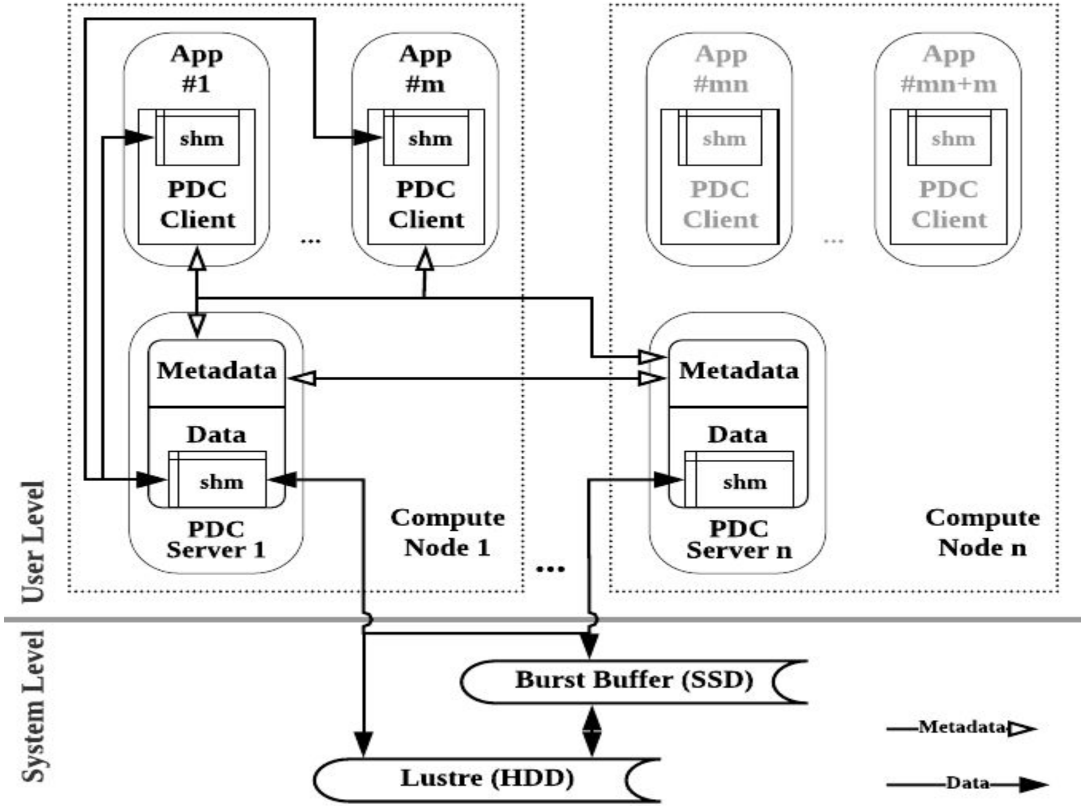
- **Scalable and asynchronous I/O**
  - Client does not stay idle to wait for I/O completion.
- **Transparent Movement between multiple storage layers.**
  - Node-local Memory/NVRAM, Burst Buffer, Lustre, etc.
- **Object-centric interface.**
  - Access data objects conveniently.
- **Direct support of multi-dimensional array and sub-region selection.**

PDC System - High-level Architecture



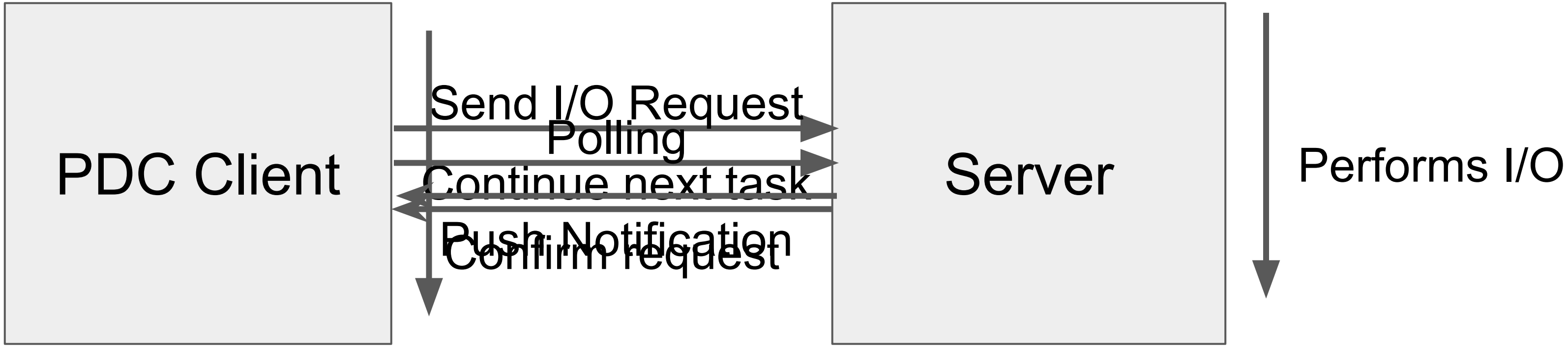


PDC System



App
#1
shm
PDC
Client
App
#m
shm
PDC
Client
App
#mn
shm
PDC
Client
App
#mn+m
shm
PDC
Client
Metadata
Data
shm
PDC
Server 1
Compute
Node 1
Metadata
Data
shm
PDC
Server n
Compute
Node n
User Level
System Level
Burst Buffer (SSD)
Lustre (HDD)
Metadata
Data

Asynchronous I/O



Storage Hierarchy-Aware Data Management

- **Memory**
  - Fastest.
  - Temporary and limited storage space.
- **Burst Buffer**
  - Fast.
  - Temporary and limited storage space.
- **Lustre**
  - Slower and requires expertise in performance tuning
  - Long term storage with enough storage space.

Data Management Optimizations

- **Node-local data aggregation**
  - Each server aggregates I/O requests from node local clients.
  - Effective use of shared memory to transfer data.
  - Log-structured write.

- **Automatic Lustre Tuning**
  - Automatically setting stripe count, size, OST index.

Metadata Optimizations

- **Collective Metadata querying.**
  - Aggregate the requests and retrieve corresponding metadata.
  - Reduce communication cost.

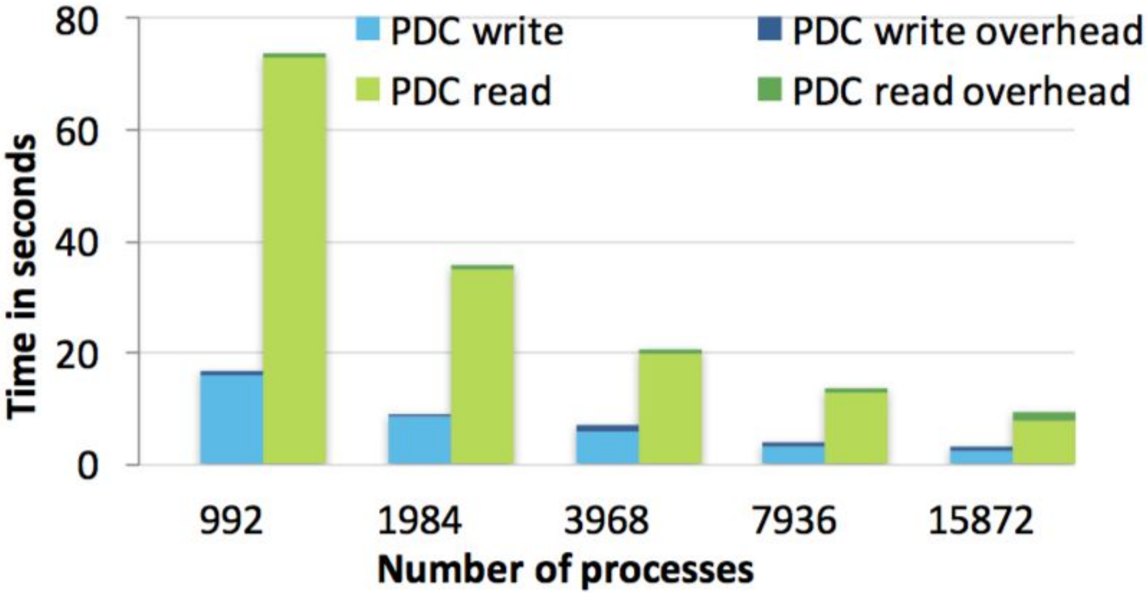
- **Relaxed metadata consistency.**
  - Delay some metadata updates and bundle with others.
  - Reduce communication cost.

Performance Evaluation

Experimental Setup

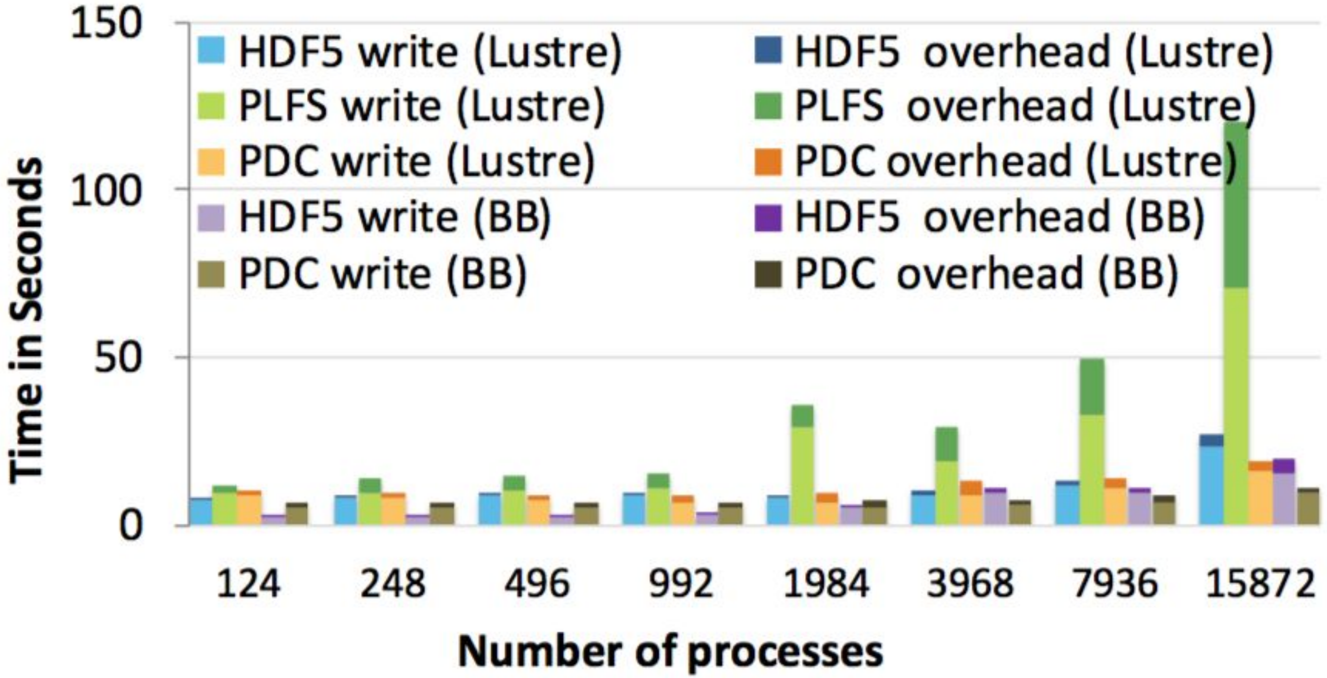
HPC Systems	Cori (NERSC), Cooley (Argonne)
Comparison	PDC, HDF5, and PLFS
Workloads	Benchmarks IO Kernels (VPIC-IO, BDCATS-IO)
Operations	Write, read with single and multiple time steps. Strong and weak scaling
Storage	Main Memory SSD-based Burst Buffer Hard disk drive (Lustre and GPFS)

I/O Strong Scaling



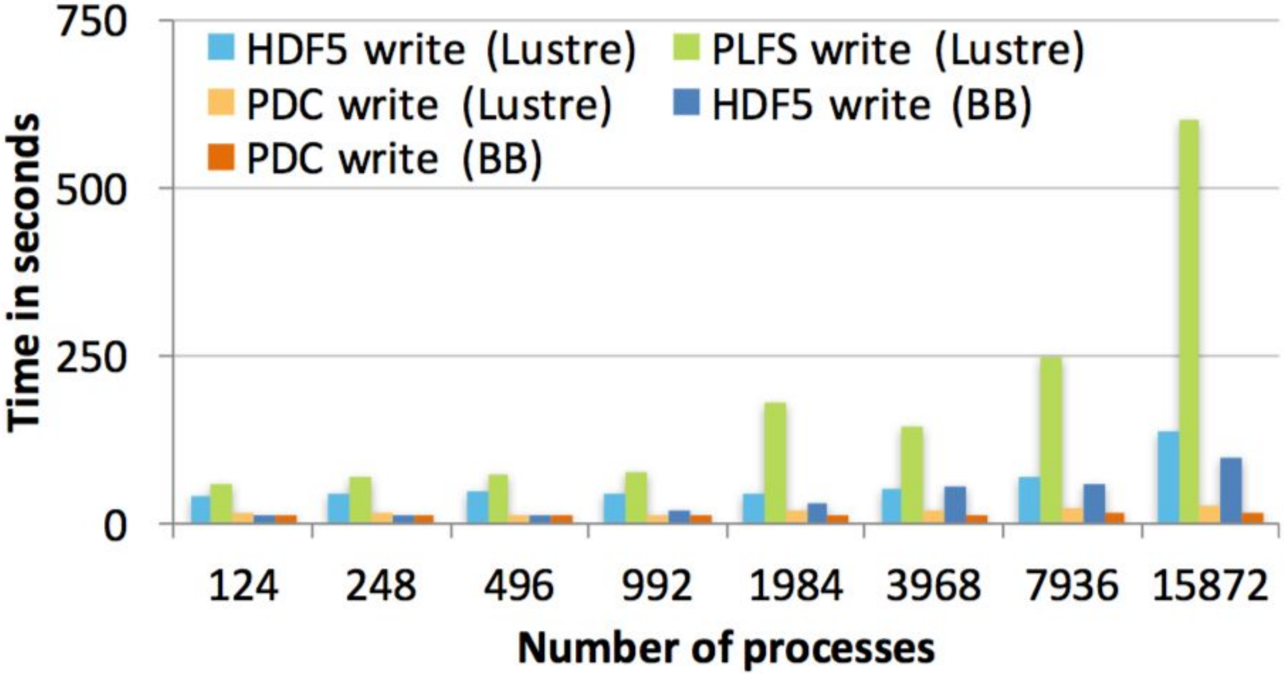
PDC strong scaling performance for writing and reading 512GB data on Lustre.

VPIC-IO (Weak Scaling) Single-timestep Write



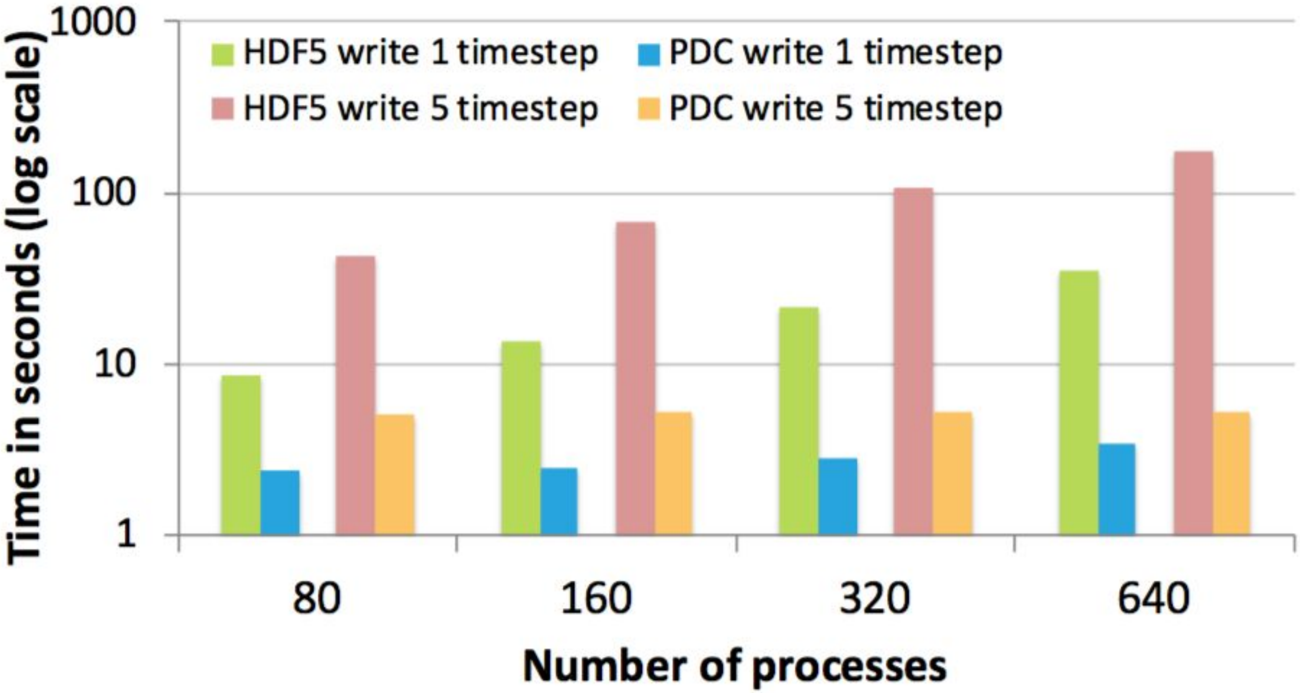
Total time for writing 1 timestep to Lustre and Burst Buffer using HDF5, PLFS, and PDC on Cori. PDC is up to **1.7x** faster than HDF5 and **9.2x** over PLFS

VPIC-IO (Weak Scaling) Multi-timestep Write



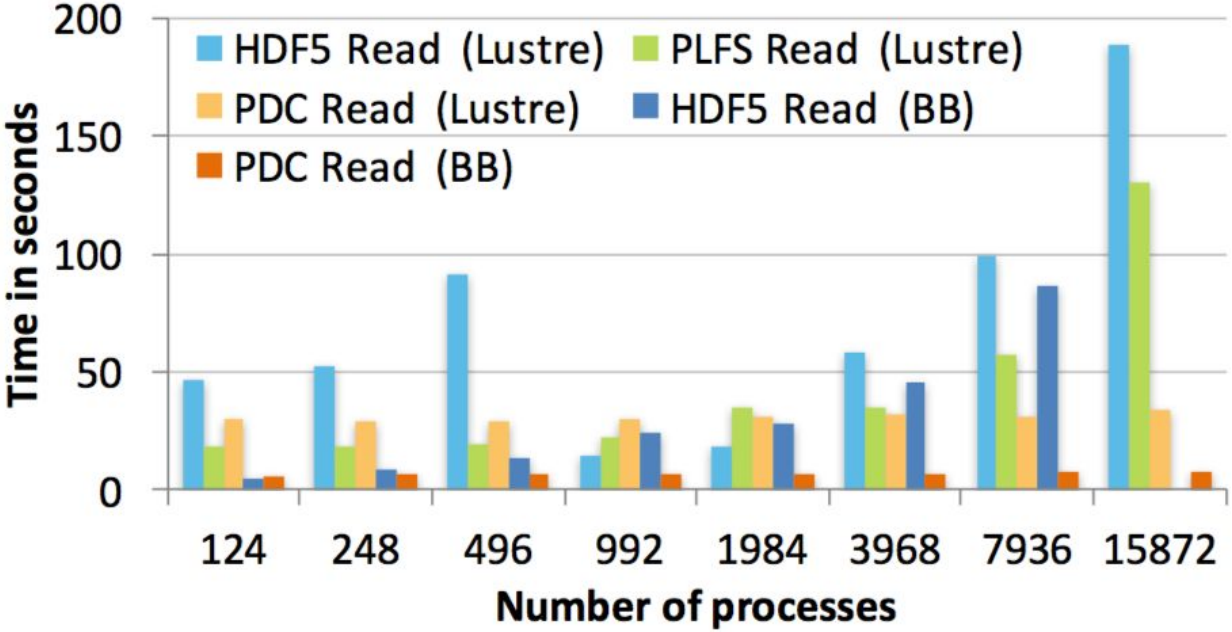
Total time to write 5 timesteps from the VPIC-IO kernel to Lustre and Burst Buffer on Cori. PDC is up to **5x** faster than HDF5 and **23x** over PLFS.

VPIC-IO Write on Cooley



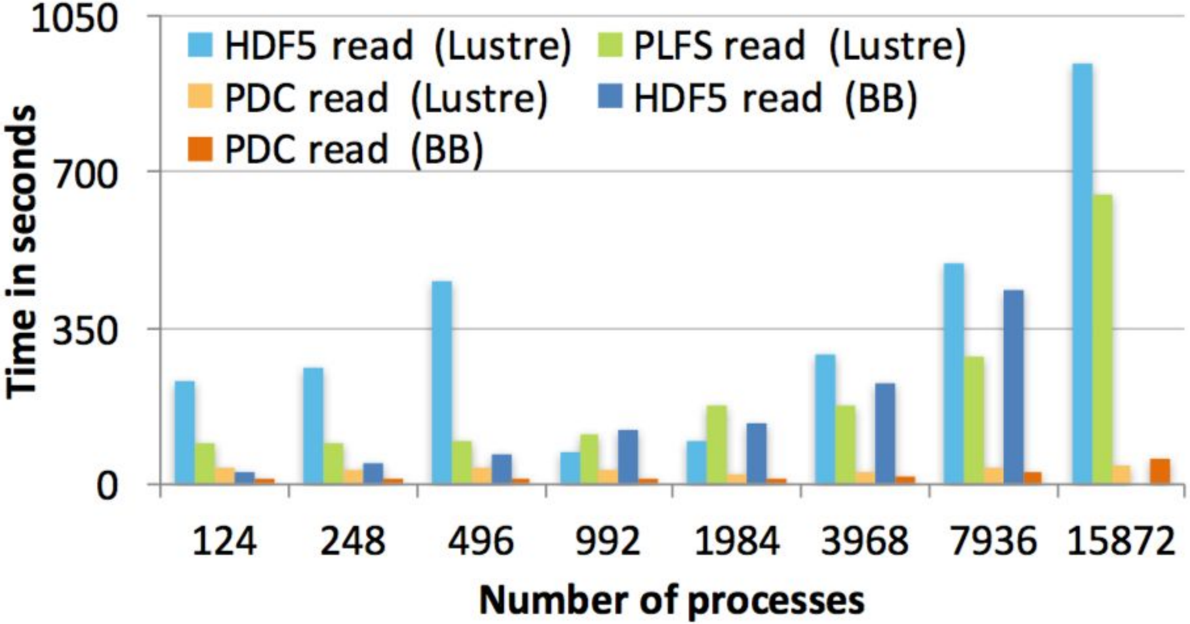
Total time to write 1 and 5 timesteps from the VPIC-IO kernel to the GPFS file system on Cooley. PDC is up to **7x** and **35x** than HDF5 to write 1 and 5 timesteps data.

BD-CATS-IO (Weak scaling) Single-timestep Read



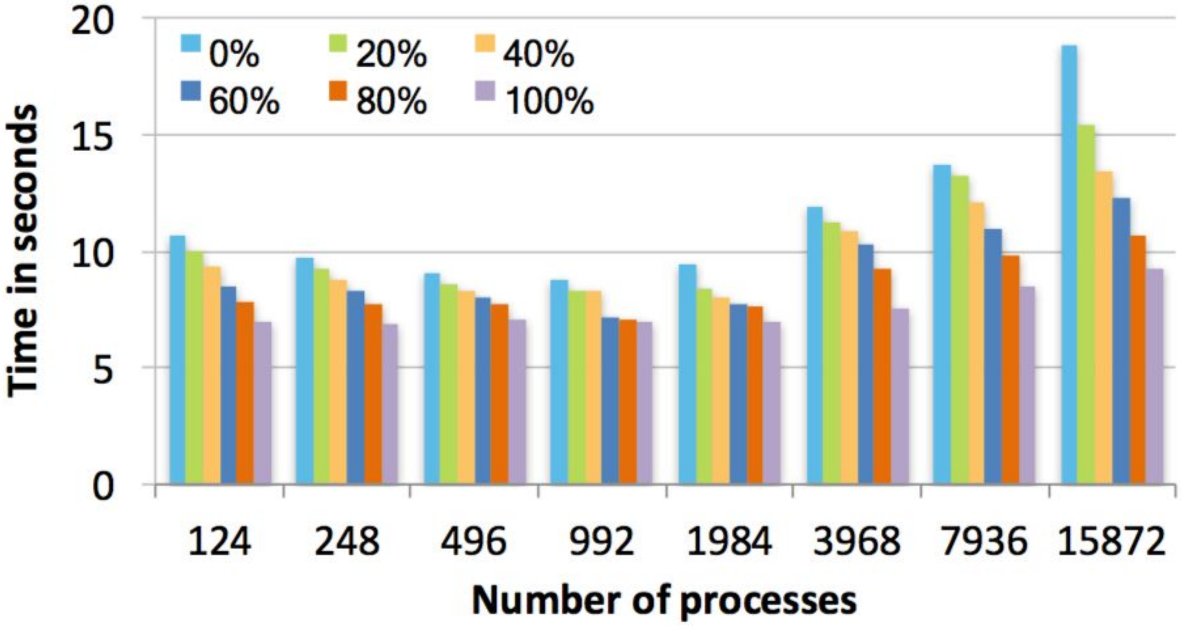
Total time for reading 1 timestep data using the BD-CATS-IO kernel using HDF5, PLFS, and PDC. PDC is up to **5x** and **4x** faster than HDF5 and PLFS.

BD-CATS-IO (Weak scaling) Multi-timestep Read



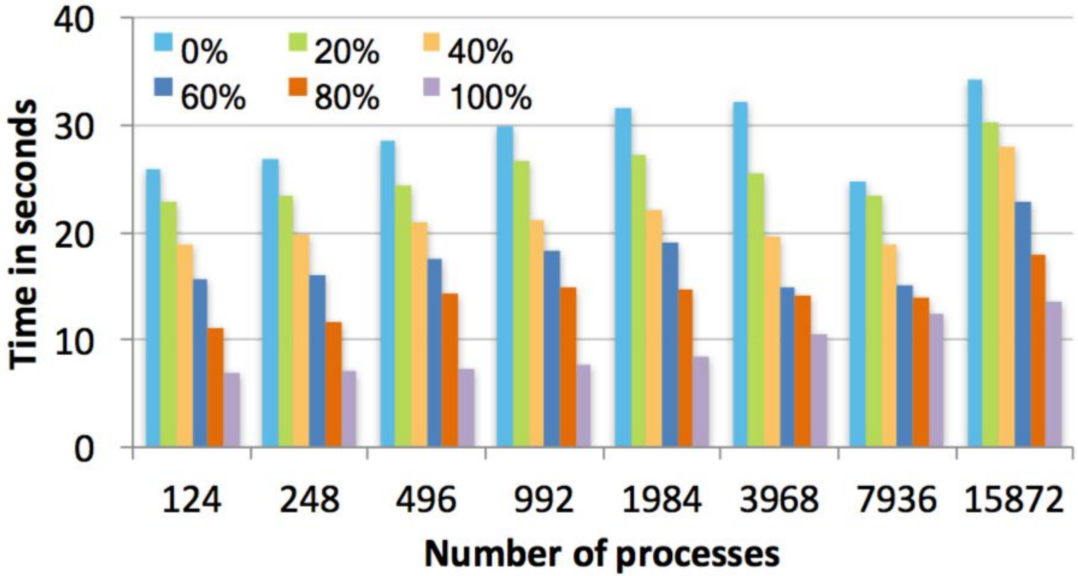
Total time for reading data of 5 timesteps from the BD-CATS-IO kernel from the Lustre and from the burst buffer. PDC is up to **11X** faster than PLFS and HDF5.

Multi-level Storage Write



Write time with part of the data written to faster burst buffer and the remaining to slower Lustre file system on Cori.

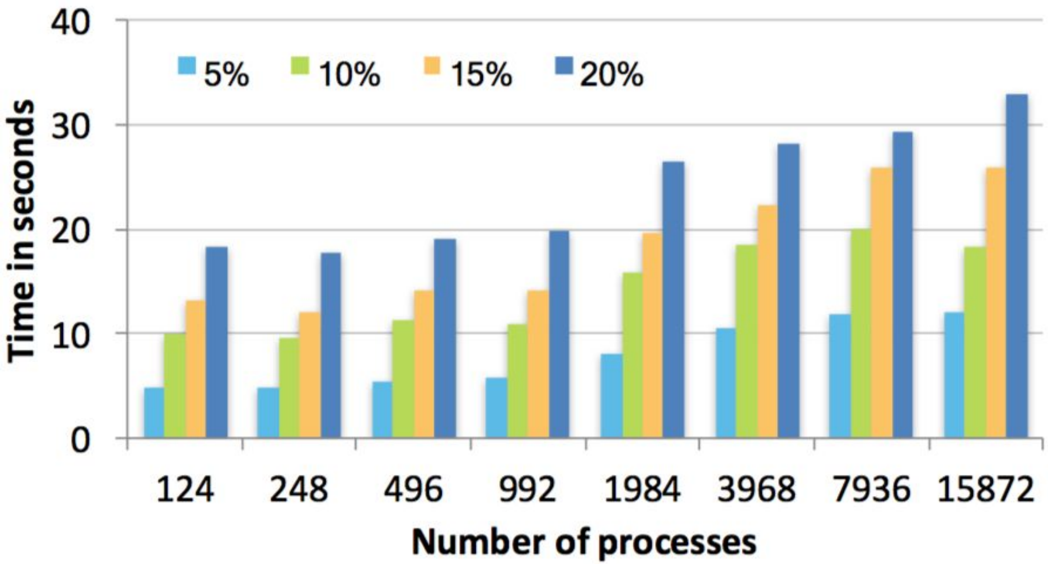
Multi-level Storage Read



Read time with part of the data written to faster burst buffer and the remaining to slower Lustre file system on Cori.

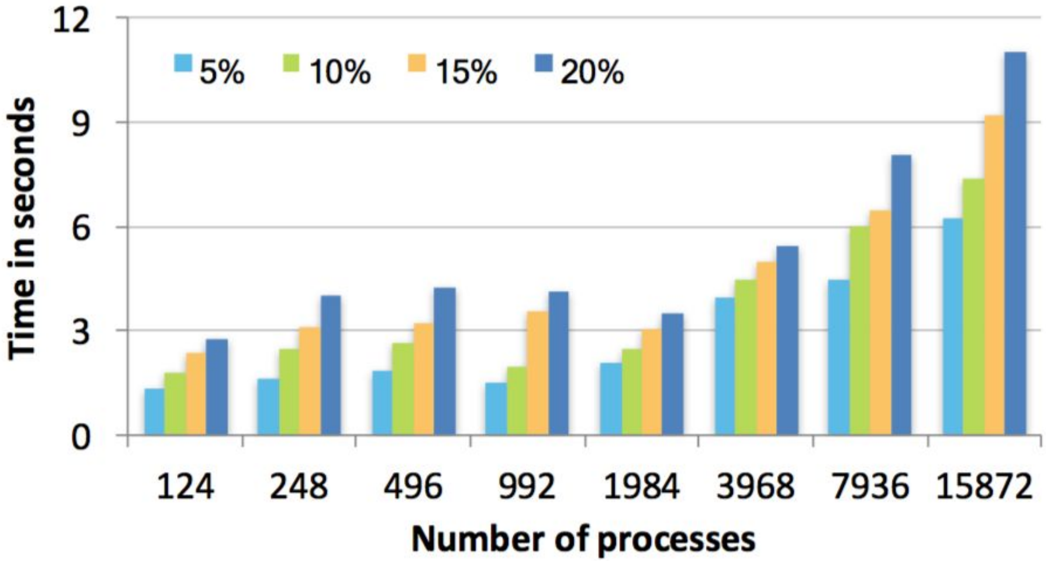
Spatial-selection Data Read from Lustre

BERKELEY LAB
Lawrence Berkeley National Laboratory



Time to read various selected object regions specified by the client processes from Lustre on Cori.

Spatial-selection Data Read from Burst Buffer



Time to read various selected object regions specified by the client processes from burst buffer on Cori.

Thanks!

Questions?

https://sdm.lbl.gov/pdc

BERKELEY LAB
Lawrence Berkeley National Laboratory