Proactive Data Containers (PDC): An Object-centric Data Store for Large-scale Computing Systems

Suren Byna

Lawrence Berkeley National Lab (LBNL), Berkeley

Co-authors

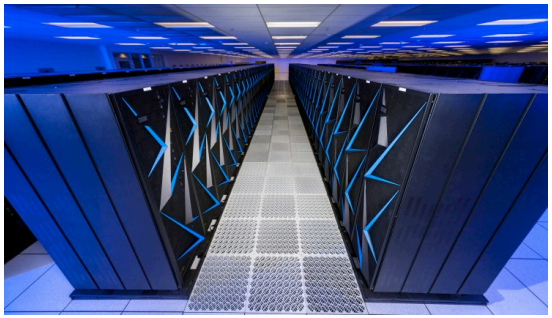
Quincey Koziol (LBNL), Venkat Vishwanath (ANL), Jerome Soumagne (THG), Houjun Tang (LBNL), Kimmy Mu (THG), Bin Dong (LBNL), Richard Warren (THG), François Tessier (ANL, now @ CSCS), Teng Wang (LBNL), and Jialin Liu (LBNL)

Scalable data management – Three disrupting trends

- Extreme parallelism
- Massive Data
- Hierarchical storage

Extreme parallelism

Summit, ORNL



Sierra, LLNL

Sunway Taihulight
NSC Wuxi, China

Summit
- ~2.4M cores
- ~143 PFlops
- 9.7 MW

Sierra
- ~1.5M cores
- ~94 PFlops
- 7.4 MW

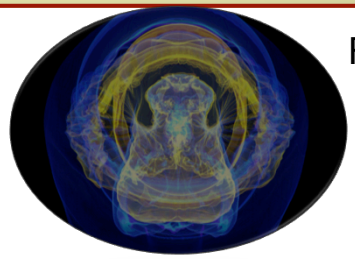
Taihulight
- ~10.6M cores
- ~93 PFlops
- 15 MW

Trinity, LANL

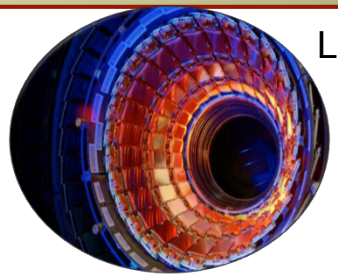
Cori, LBNL

Massive scientific data

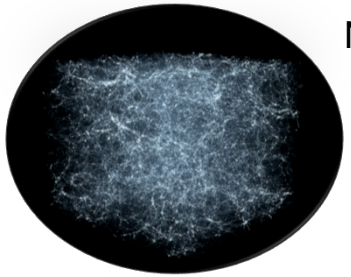
- Simulations
  - Multi-physics (FLASH) – 10 PB
  - Cosmology (NyX) – 10 PB
  - Plasma physics (VPIC) – 1 PB
- Experimental and observational data (EOD)
  - LHC (100 PB),
  - LSST (60 PB),
  - Genomics (100 TB to 1 PB)



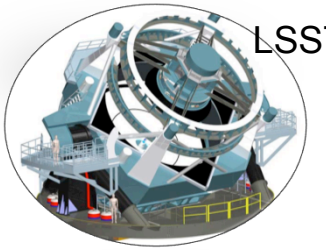
FLASH



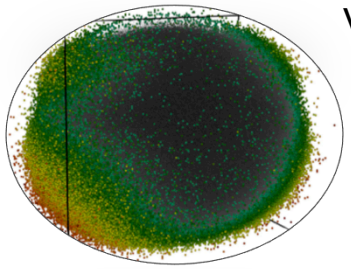
LHC



NyX



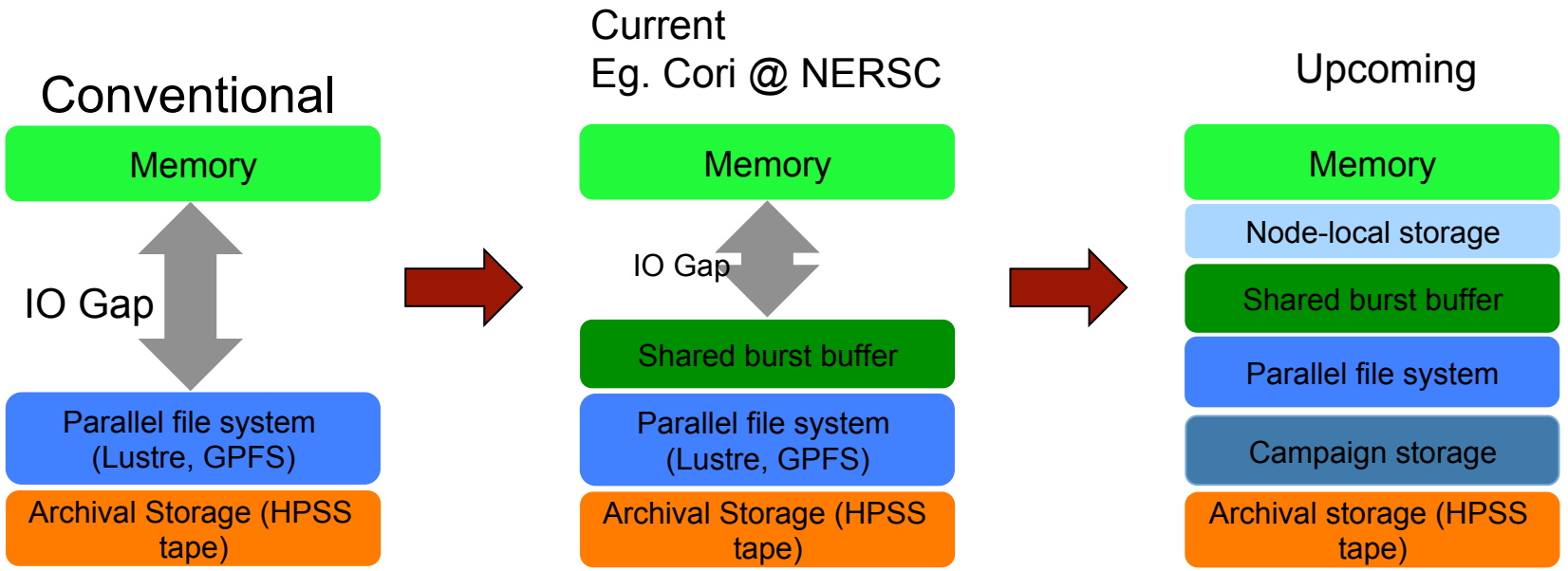
LSST



VPIC

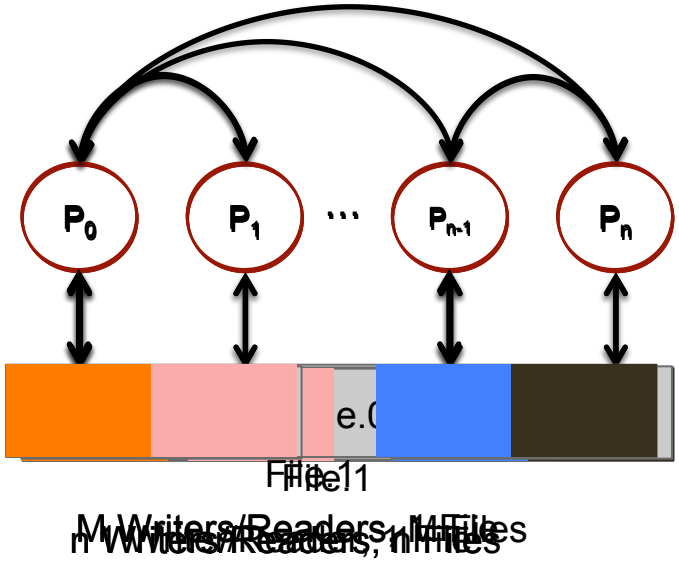
Genomics

Hierarchical and heterogeneous storage

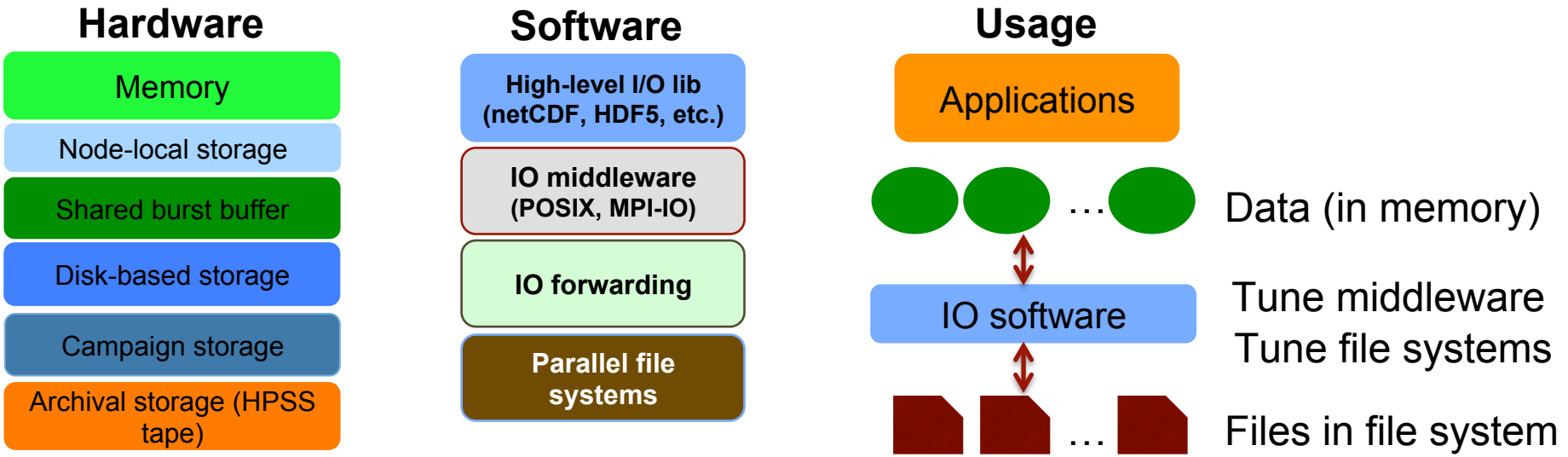


Reading and writing data on scalable systems

- **Types of parallel I/O**
  - 1 writer/reader, 1 file
  - N writers/readers, N files (File-per-process)
  - N writers/readers, 1 file
  - M writers/readers, 1 file
    - Aggregators
    - Two-phase I/O
  - M aggregators, M files (file-per-aggregator)
    - Variations of this mode



Scalable Storage Systems: Challenges



- **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**

Scalable data management requirements

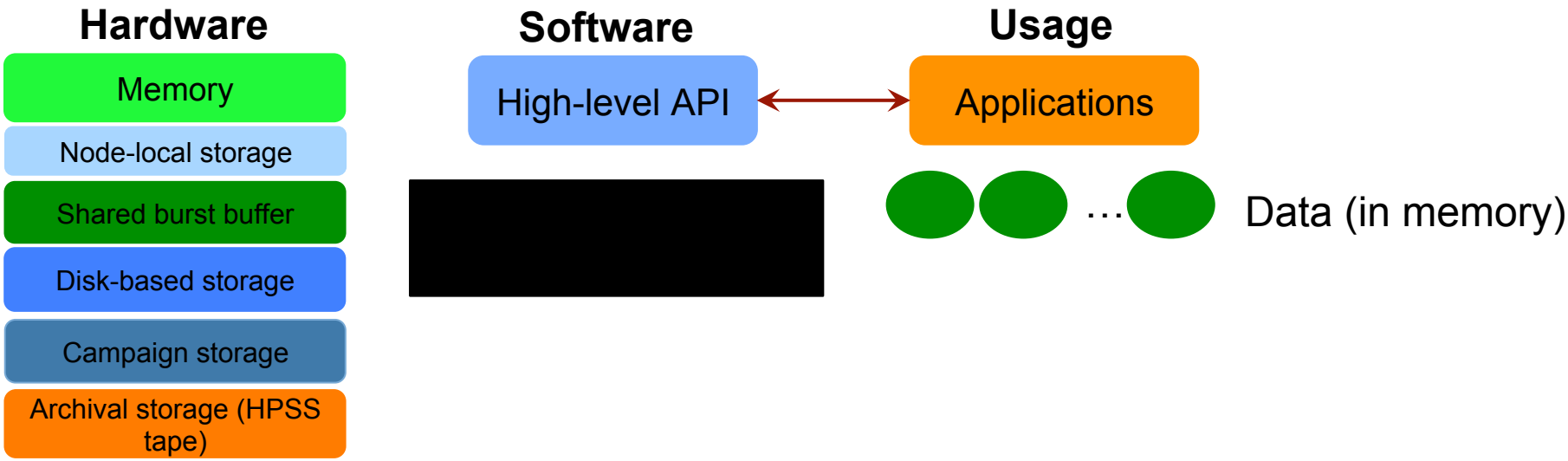
Use case	Domain	Sim/EOD/ analysis	Data size	I/O Requirements
FLASH	High-energy density physics	Simulation	~1PB	Data transformations, scalable I/O interfaces, correlation among simulation
CMB / Planck	Cosmology	Simulation, EOD/	10PB	Automatic data movement optimizations
DECam & LSST				ata transformations
E3SM				atic
TECA				movement
HipMer	Genomics	EOD/Analysis	~100TB	Scalable I/O interfaces, efficient and automatic data movement

Easy interfaces and superior performance

Transparent data management

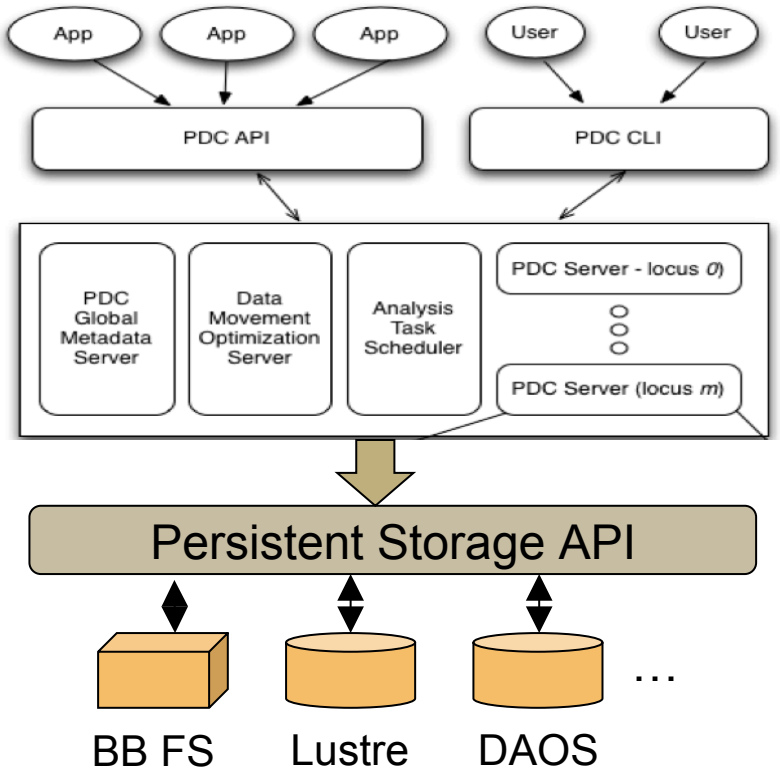
Information capture and management

Next Gen Storage – Proactive Data Containers (PDC)



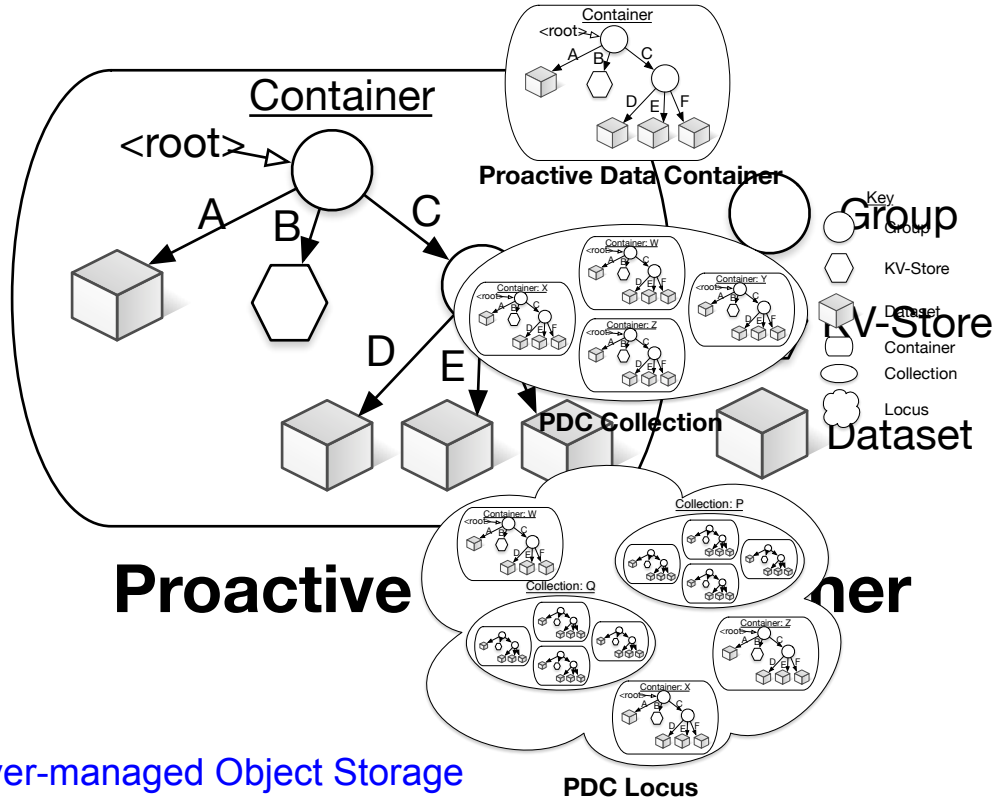
PDC System – High-level Architecture

- **Object-centric data access interface**
  - Simple put, get interface
  - Array-based variable access
- **Transparent data management**
  - Data placement in storage hierarchy
  - Automatic data movement
- **Information capture and management**
  - Rich metadata
  - Connection of results and raw data with relationships



Object-centric PDC Interface

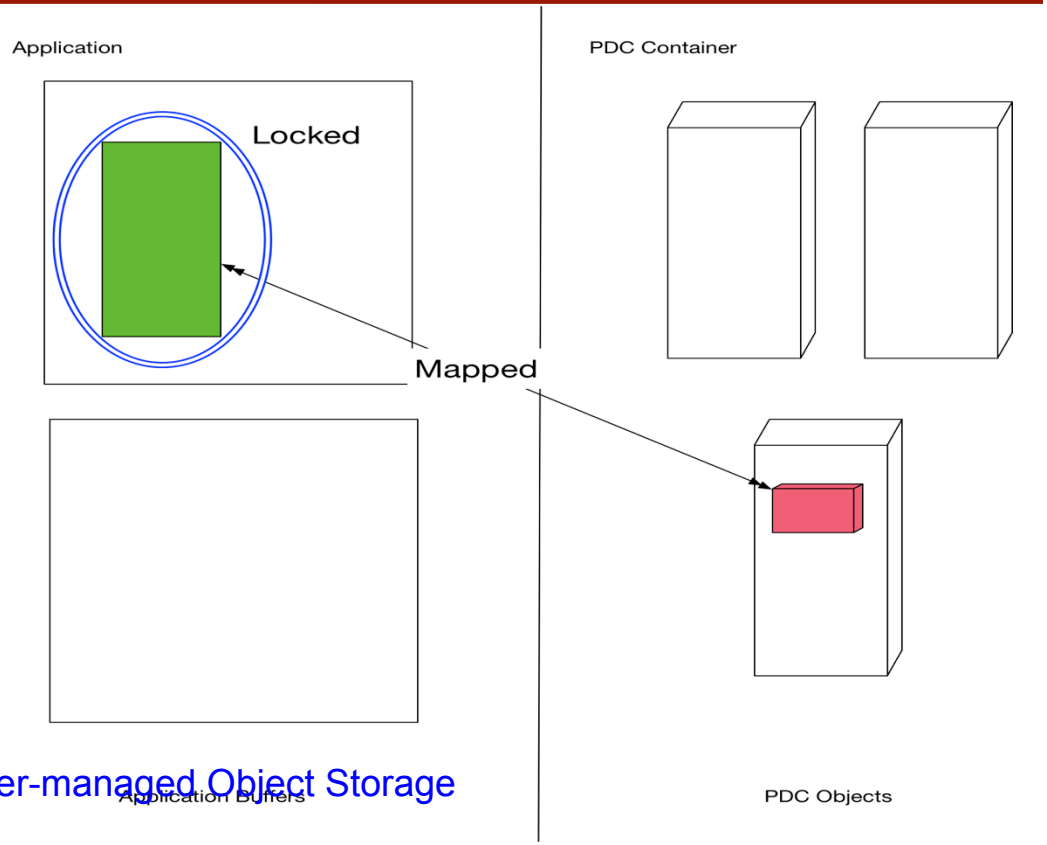
- Object-level interface
  - Create – containers and objects
  - Add attributes
  - Put object
  - Get object
  - Delete object
- Array-specific interface
  - Create regions
  - Map regions in PDC objects
  - Lock
  - Release



J. Mu, J. Soumagne, et al., "A Transparent Server-managed Object Storage System for HPC", IEEE Cluster 2018

Object-centric PDC Interface

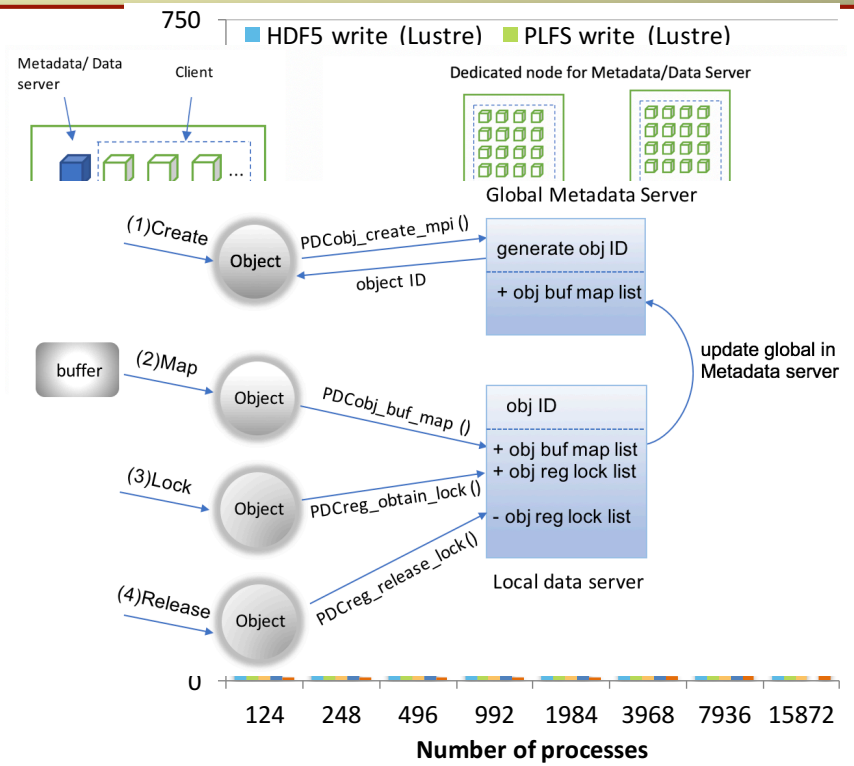
- Object-level interface
  - Create – containers and objects
  - Add attributes
  - Put object
  - Get object
  - Delete object
- Array-specific interface
  - Create regions
  - Map regions in PDC objects
  - Lock
  - Release



J. Mu, J. Soumagne, et al., "A Transparent Server-managed Object Storage System for HPC", IEEE Cluster 2018

Transparent data movement in storage hierarchy

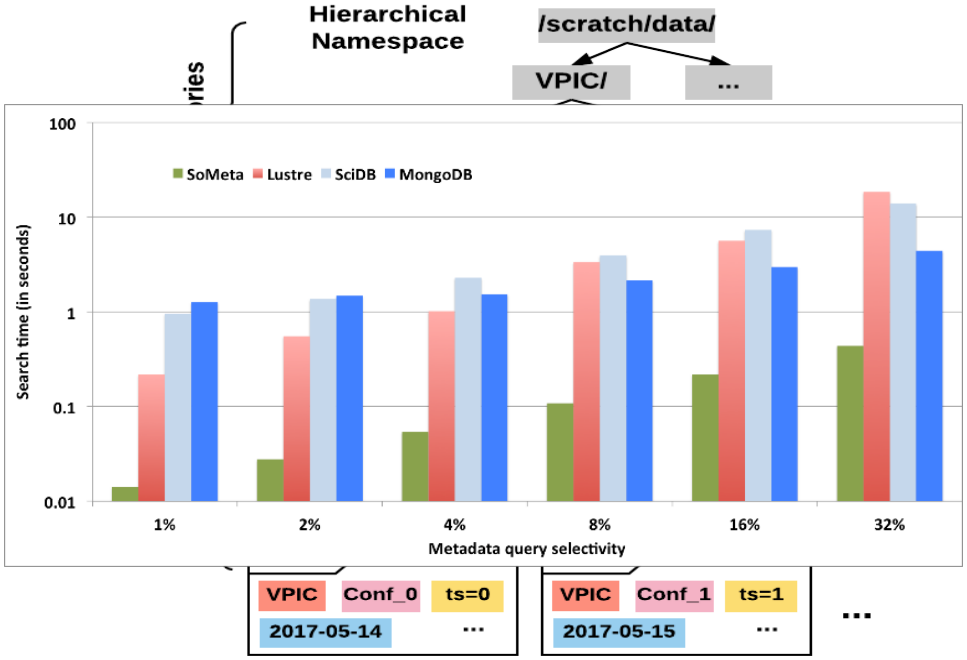
- Usage of compute resources for I/O
  - Shared mode – Compute nodes are shared between applications and I/O services
  - Dedicated mode – I/O services on separate nodes
- Transparent data movement by PDC servers
  - Apps map data buffers to objects and PDC servers place and manage data
  - Apps query for data objects using attributes
- Superior I/O performance



H. Tang, S. Byna, et al., "Toward Scalable and Asynchronous Object-centric Data Management for HPC", IEEE/ACM CCGrid 2018

Metadata management

- Flat name space
- Rich metadata
  - Pre-defined tags that includes provenance
  - User-defined tags for capturing relationships between data objects
- Distributed in memory metadata management
  - Distributed hash table and bloom filters used for faster access



H. Tang, S. Byna, et al., "SoMeta: Scalable Object-centric Metadata Management for High Performance Computing", to be presented at IEEE Cluster 2017

Conclusions

- Take home message
  - **Scalable storage systems impacted by:**
    - Extreme level of parallelism
    - Massive amounts of scientific data
    - Transforming storage architectures
  - **Proactive data containers**
    - Object-centric interfaces
    - Transparent data movement in storage hierarchies
    - Scalable management of extensive metadata

Thanks

https://sdm.lbl.gov/pdc

Contact: Suren Byna (SByna@lbl.gov)