Byna S, Breitenfeld M S, Dong B et al. ExaHDF5: Delivering efficient parallel I/O on exascale computing systems. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 35(1): 145–160 Jan. 2020. DOI 10.1007/s11390-020-9822-9

ExaHDF5: Delivering Efficient Parallel I/O on Exascale Computing Systems

EXASCALE COMPUTING PROJECT

Suren Byna (Lawrence Berkeley Lab - LBL)

Scot Breitenfeld (The HDF Group - THG), Bin Dong (LBL), Quincey Koziol (LBL), Elena Pourmal (THG), Dana Robinson (THG), Jerome Soumagne (THG), Houjun Tang (LBL), Venkatram Vishwanath (Argonne National Lab), and Richard Warren (THG)

BERKELEY LAB
Lawrence Berkeley National Laboratory

Argonne NATIONAL LABORATORY

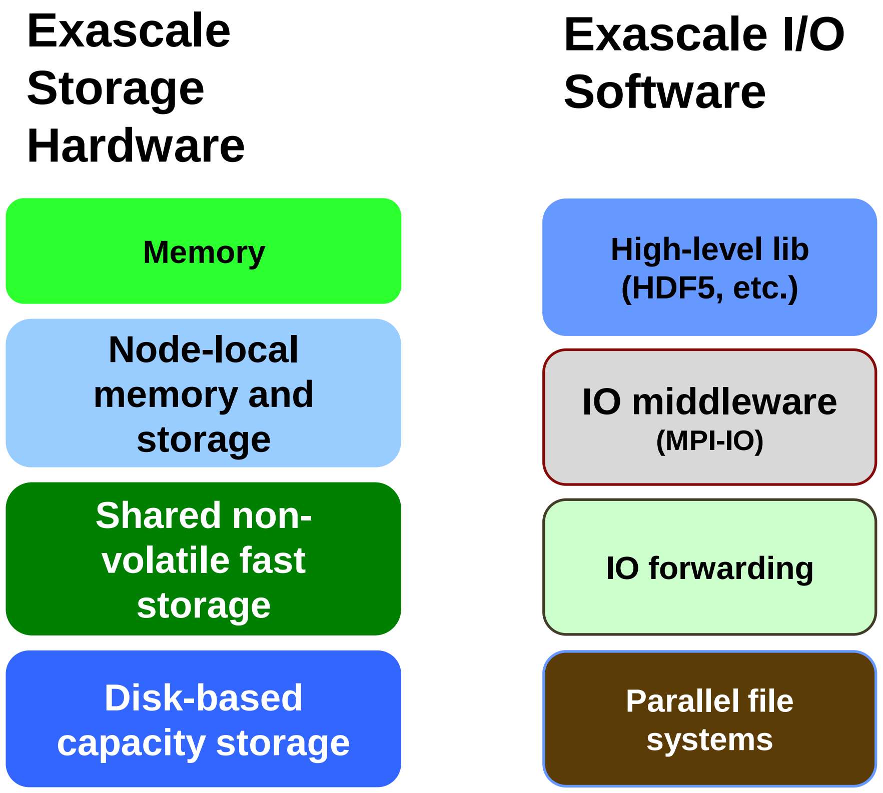
The HDF Group

NNSA National Nuclear Security Administration

U.S. DEPARTMENT OF ENERGY | Office of Science

Exascale I/O architectures and software

- Exascale storage hardware
  - Deepening hierarchy with:
    - Fast node-local storage and storage-class memory
    - Shared SSD-based storage layer
    - Disk-based capacity storage
- I/O software
  - High-level self-describing I/O libraries (HDF5, etc.)
  - Middleware (MPI-IO) and optimization layers
  - File systems (Lustre and Spectrum Scale/GPFS) and object storage (Intel DAOS)
- Challenges
  - Heterogeneity of storage devices and distributed across nodes
  - Disparity of I/O software stack (different tuning parameters)
  - Overheads of managing metadata in self-describing formats
  - Obtaining sustained I/O performance on exascale storage



Exascale Storage Hardware

Memory

Node-local memory and storage

Shared non-volatile fast storage

Disk-based capacity storage

Exascale I/O Software

High-level lib (HDF5, etc.)

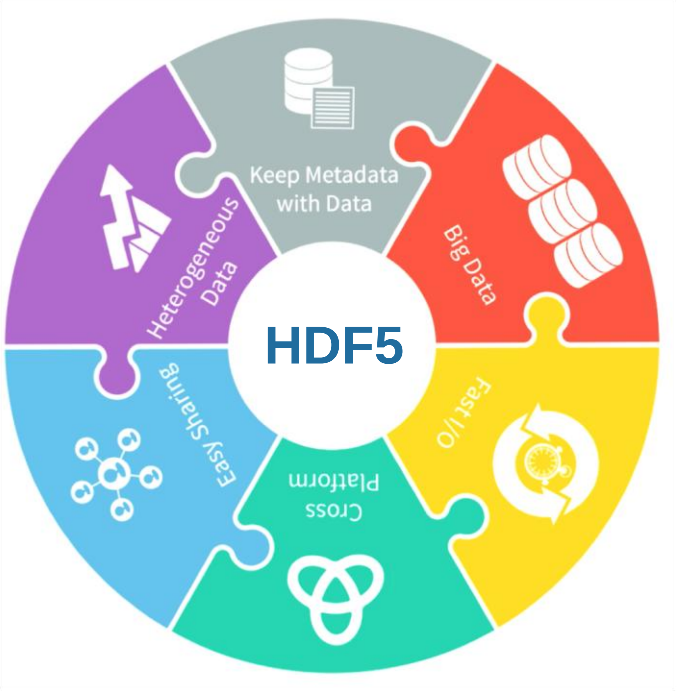
IO middleware (MPI-IO)

IO forwarding

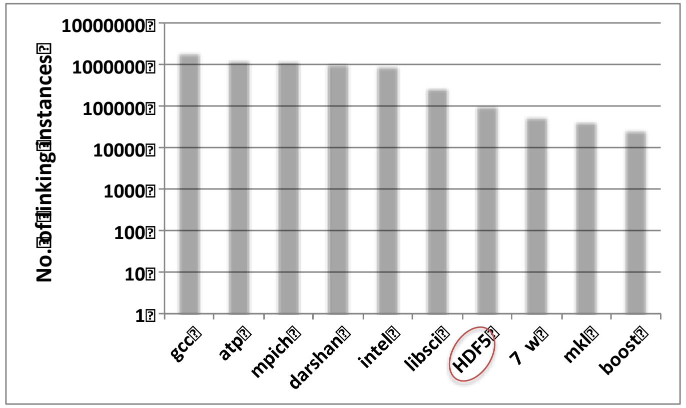
Parallel file systems

ExaIO Project Products – HDF5

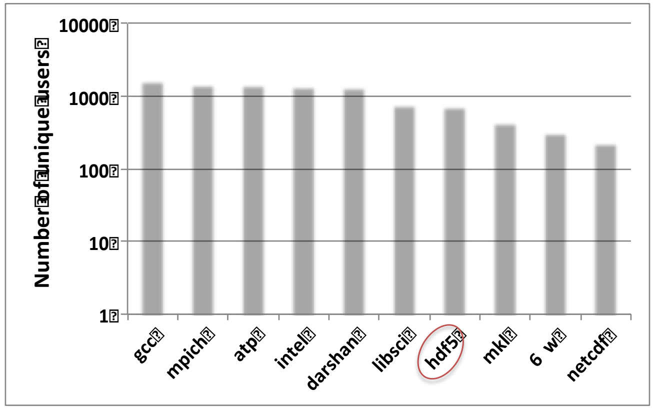
- HDF5 is designed to organize, store, discover, access, analyze, share, and preserve diverse, complex data in continuously evolving heterogeneous computing and storage environments.
  - Maintained by The HDF Group (THG)
- NASA/NOAA satellite data (Aura, JPSS-1, etc.)
  - Highest Technology Readiness Level (TRL 9) - “Flight proven” through successful mission operations
- Heavily used on DOE supercomputing systems



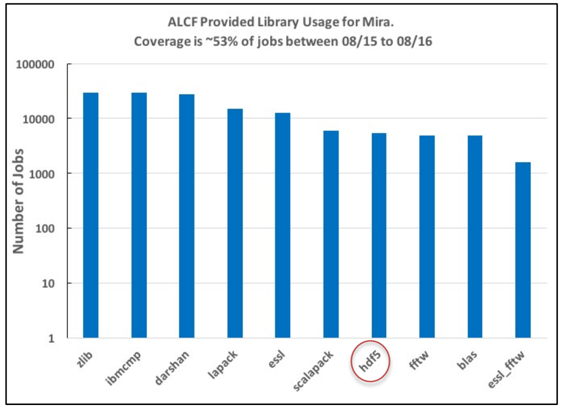
HDF5: 2002 R&D 100 Award Winner



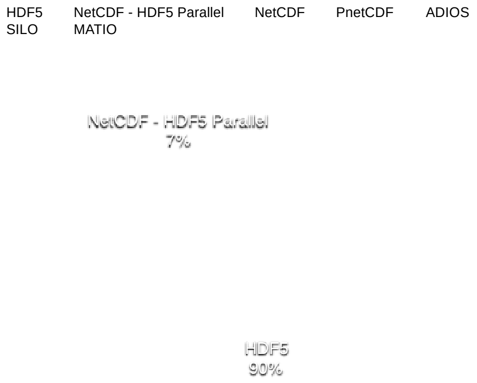
a. Number of linking instances on Edison (NERSC)



b. Number of unique users on Edison (NERSC)

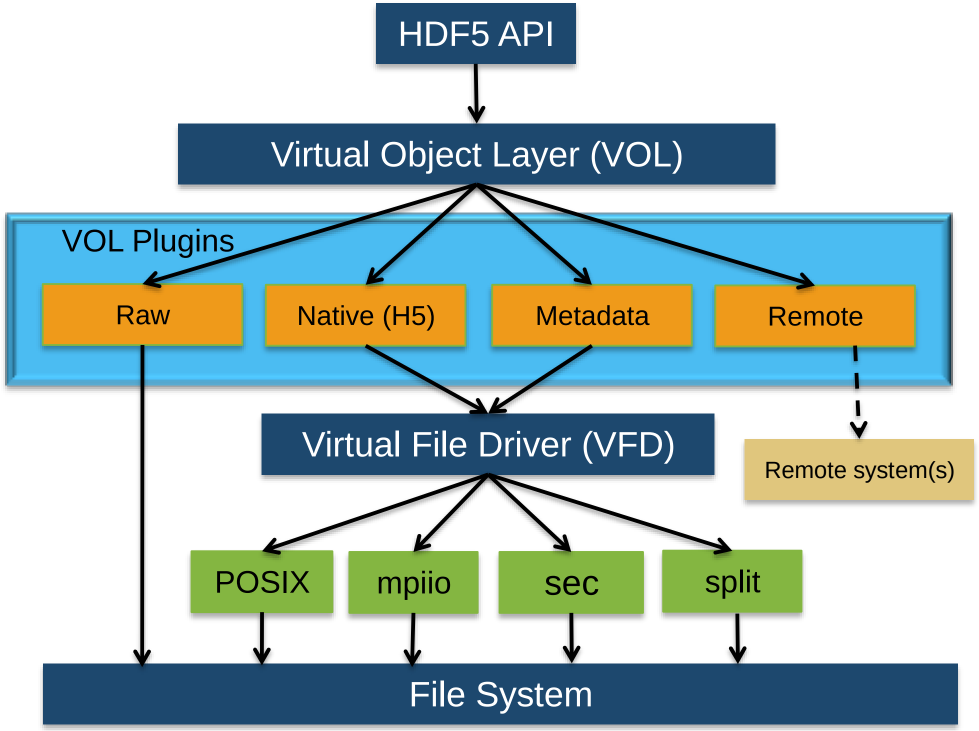


c. Number of linking instances on Mira (ALCF)



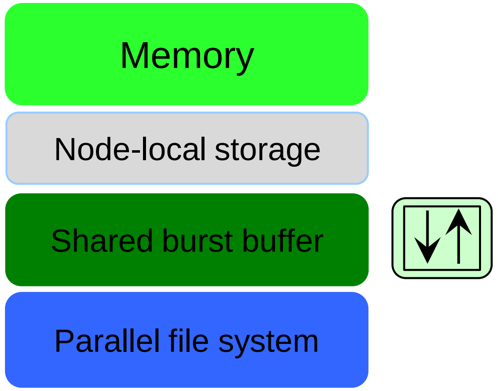
Virtual Object Layer (VOL)

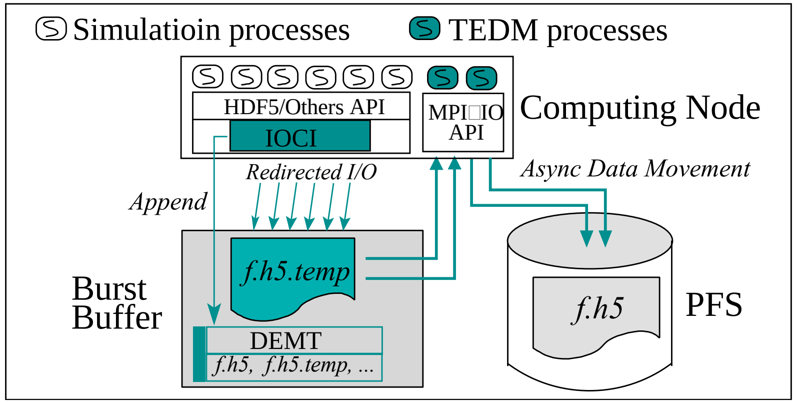
- Virtual Object Layer (VOL) provides an application with the HDF5 data model and API, but allow different underlying storage mechanisms
- Enables developers to use HDF5 on novel current and future storage systems easily
  - Prototype VOL connectors for using burst buffer storage transparently and for accessing DAOS are available
  - Developed VOL connectors for reading PnetCDF and ADIOS-BP data
- Integrated into the HDF5 trunk (will be released in 1.10.12 later this year) https://bitbucket.hdfgroup.org/projects/HDFFV/repos/hdf5/
- Allows ADIOS and other libraries to use HDF5 API

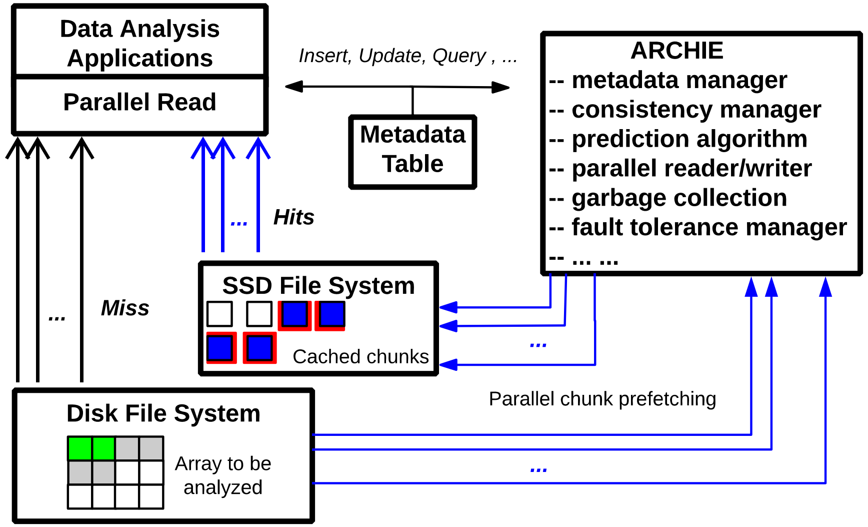


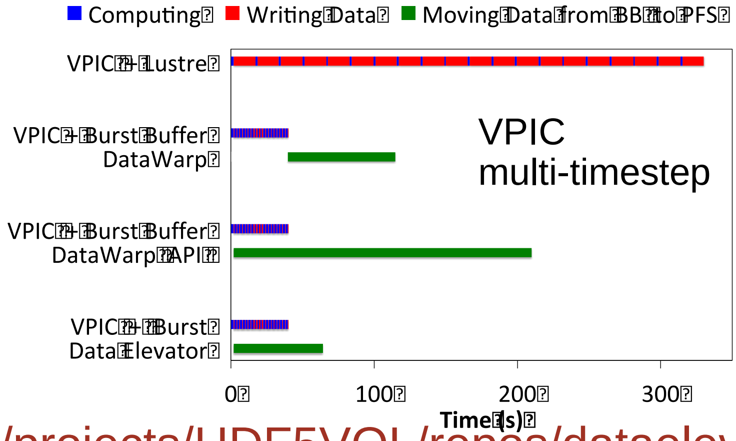
HDF5 Data Elevator

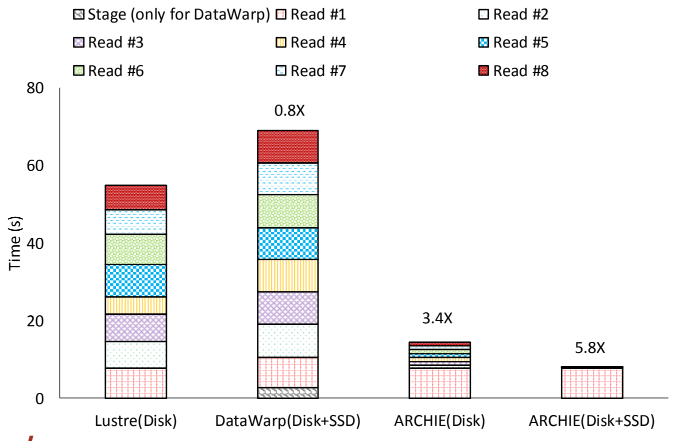
- Data Elevator VOL connector
  - Transparent data movement in storage hierarchy - writes and reads
  - Intercepts file opens, write, read, and close function calls and places data in burst buffers temporarily; DE moves data asynchronously
  - Prefetches predicted chunks of data to burst buffer or memory
  - In situ data analysis capability using burst buffers
  - Phase 2 plan includes extending capabilities of Data Elevator for node-local storage







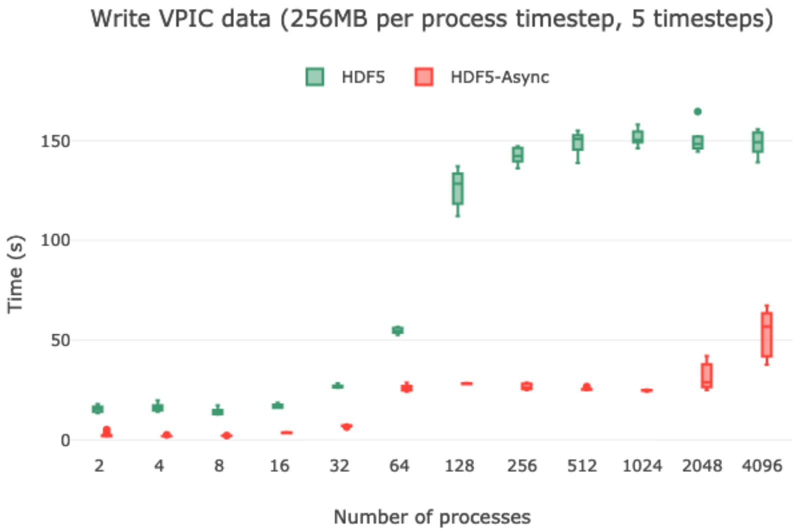


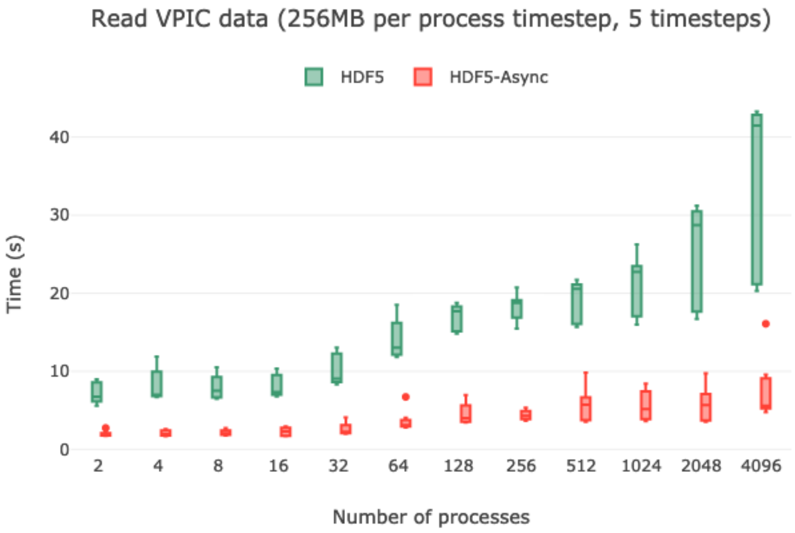


https://bitbucket.hdfgroup.org/projects/HDF5VOL/repos/dataelevator/

Asynchronous I/O with HDF5

- Asynchronous I/O allows an application to overlap I/O with other operations
- The asynchronous I/O feature has been implemented as a VOL (Virtual Object Layer) connector, without requiring major change the HDF5 library





EXASCALE COMPUTING PROJECT

https://bitbucket.hdfgroup.org/projects/HDF5VOL/repos/async/