HDF5 Roadmap for 2019-2020 and Beyond

September 17, 2019

Elena Pourmal epourmal@hdfgroup.org

The HDF Group

Outline

- **HDF5 Open Source Project**
- **HDF5 Challenges**
- **HDF5 Major Capabilities**
- **HDF5 Roadmap for 2019-2020**
  - HDF5 releases schedule
- **New Features in Design**
- **SWMR Demo**

HDF5 Open Source Project

Challenges and solutions

HDF5 Open Source Project Challenges

- **OSS Project = Software + Community**
  - Project is not sustainable without community support and involvement
- **HDF5 project lacks community contributions**
  - Complex and undocumented software architecture
  - Obscure location of source code repository, issues database and documentation
  - Cumbersome process of submitting patches
  - Lack of coding and development process standards
  - Minimum time available to The HDF Group developers to work with the contributors
  - "Conflict" between community interests and The HDF Group's paid projects
  - ….

HDF5 Open Source Project Challenges (cont'd)

- **How can we solve this porblem?**
  - Address the issues of HDF5 developers' documentation
  - Pubish our internal standards
  - Grow a team of external developers by working with individulas
  - Collaborate with developers of HDF5 ecosystem software (e.g., h5py, HDFql) and encourage them to contribute to HDF5
  - Outreach:
    ‣ Workshops
    ‣ Users' groups
    ‣ Technical Webinars
    ‣ Blogs

Building HDF5 Contributors Community

- **HDF5 JIRA access**
  - HDF5 JIRA is open
    - ‣ Register at https://hdfgroup.org
    - ‣ Login to https://jira.hdfgroup.org with your login and password
      - ‣ Browse issues
      - ‣ Vote for issues
- **HDF5 Help Desk**
  - Submit and watch your issues via Atlassian Service Desk
- **Access to HDF5 Bitbucket**
  - We opened access to HDF5 repository in Bitbucket
    - ‣ Register at https://hdfgroup.org
    - ‣ Contact help@hdfgroup.org for Bitbucket account
- **Contact us to present your project at our Webinar or write a blog**
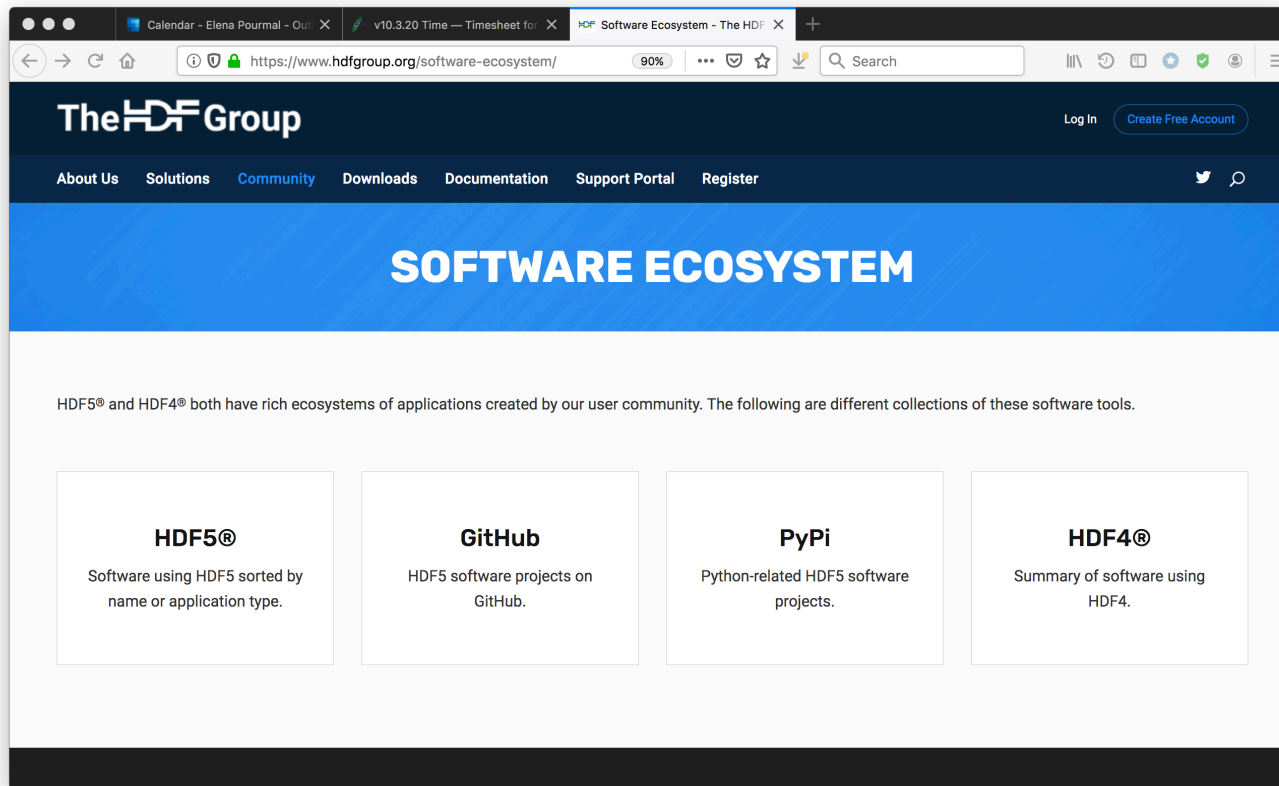 - HDFql talk will be scheduled in November/December 2019

Strengthening Connections with HDF5 Ecosystem Community

- **Community Portal** **https://www.hdfgroup.org/community/**

Connecting to HDF5 Ecosystem Community

- **Github has 1,533 HDF5 related projects**
  - Language wrappers or High-Level APIs
  - Converters (PyROOT, Pandas-to-Postgress)
  - Visualization and browsing tools
  - Packages to work with specific types of data (e.g., times series)
  - Focus on ease of access to data stored in HDF5
  - Many are dormant or abandoned
  - Let use know about your software and we will add it to our Community Portal
  - Present at our Webinar and internal technical seminar
- **Invite us to collaborate**
  - Proposals
  - Papers
  - Conferences, etc.

h5py collaboration

- **Current collaboration with h5py developers**
  - Minimum interaction and contributions
  - The HDF Group members are working on code contribution for new chunk query functions
    ‣ Released in 1.10.5 and 1.12.0
    ‣ Provide locations, lengths and filters information
    ‣ Can be used for direct access to data stored in HDF5 without HDF5 library
  - Monitor and answer questions on h5py FORUM (Github)
  - Use h5py and Kita Lab in The HDF Group Tutorials
- **Plans**
  - Work with h5py community on CI with HDF5 under development
  - Engage h5py representative(s)

HDFql – High Level Language to Manage HDF5 Data

HDFql

Simple
Designed to be as simple as SQL. Hides complex operations and frees users from low-level details

Clean
Offers a clean interface requiring just a few intuitive statements even for complex operations. Gone are the days where HDF5 required endless lines of code

Fast
Processes data using all nodes and cores available. This means much higher volumes of data are processed in the same amount of time

Intuitive
Based on models of human cognition and natural language. Fast learning curve

Reliable
Unfailing robustness and reliability. HDFql is checked against hundreds of existing tests before new versions are released

Writes
Unlike other tools, HDFql not only reads HDF5 but also allows you to write HDF5 data

Portable
Portable across C, C++, Java, Python, C#, Fortran and R using one uniform high-level language

Platforms
Supports Windows, Linux, macOS

HDFql collaboration

The HDF Group

- **Presented at ISC 19 in Frankfurt, Germany**
- **Scheduled to present at THG Technical Seminar later this Fall**
- **HDFql Blog**
  - https://www.hdfgroup.org/2016/06/hdfql-new-hdf-tool-speaks-sql/
  - New blog is coming!

HDF5 Challenges

Performance

Extending HDF5 capabilities

HDF5 Challenges

- Maintenance of core HDF5 library and tools
- Modularity for easy maintenance and external contributions
- Architectural and maintenance documentation
- Backward and forward compatibility
- Performance
  - Establish performance regression testing
  - Enhance sequential performance
    ‣ Hyperslab selection mechanism
    ‣ Support for many (>1K) objects in the file

HDF5 Performance
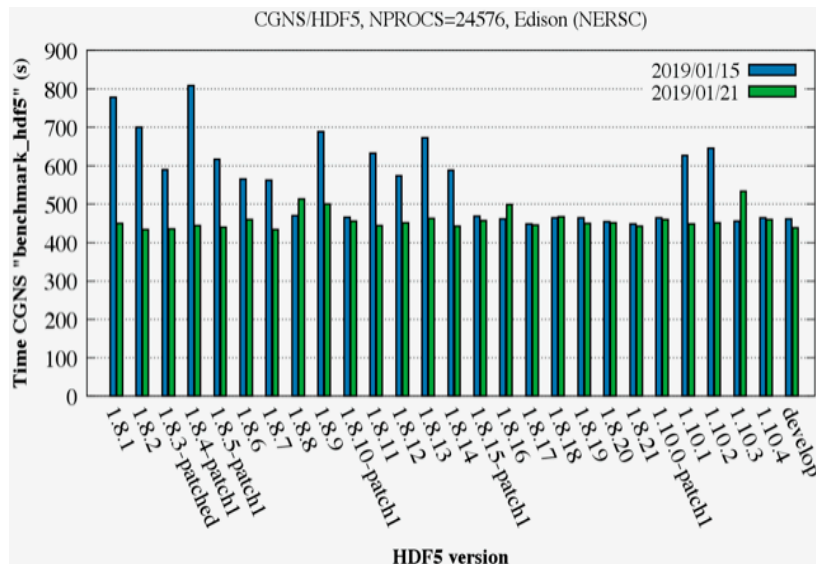
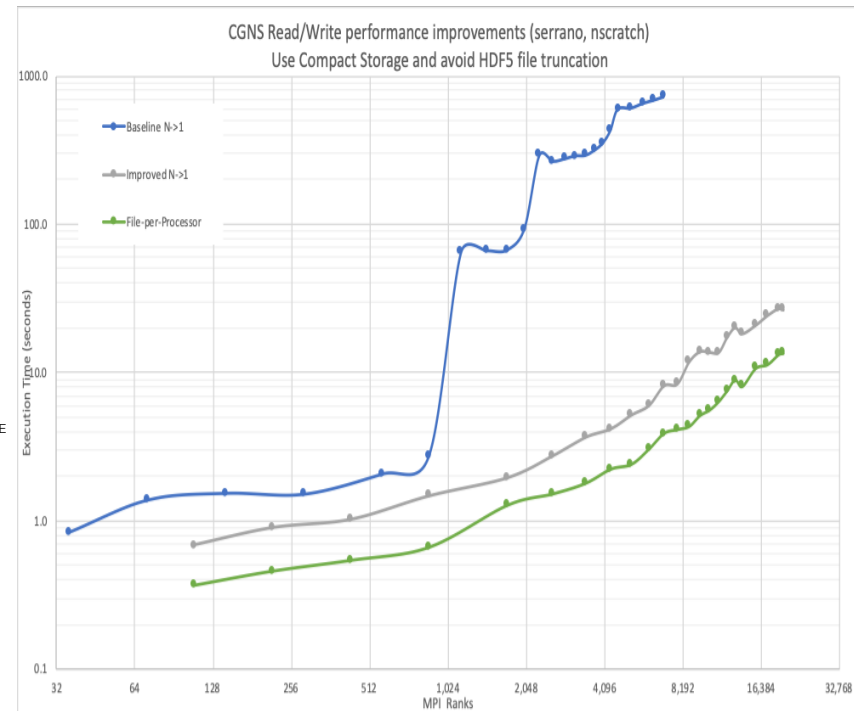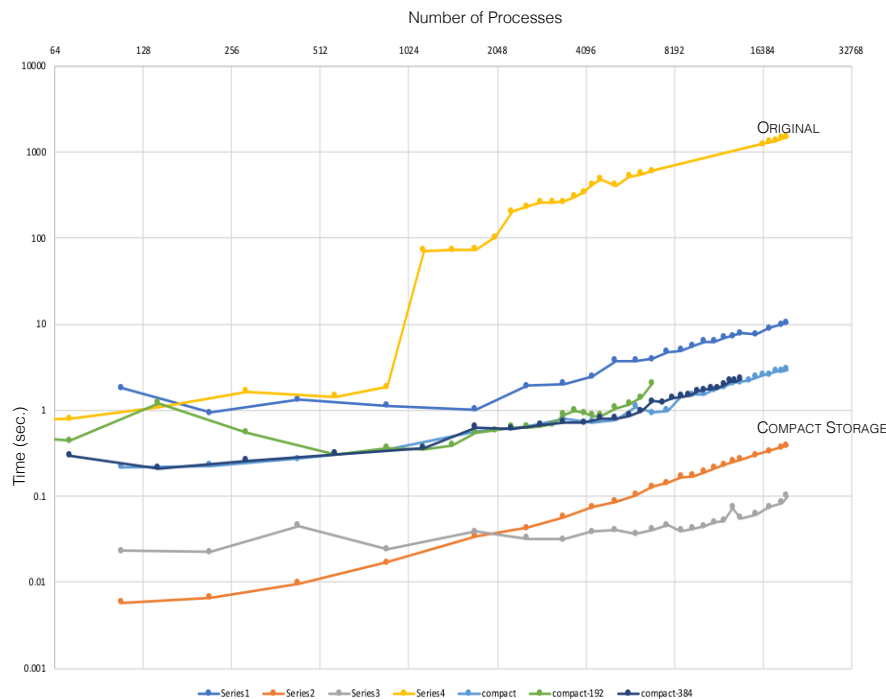CGNS library: Testing times with different HDF5 versions
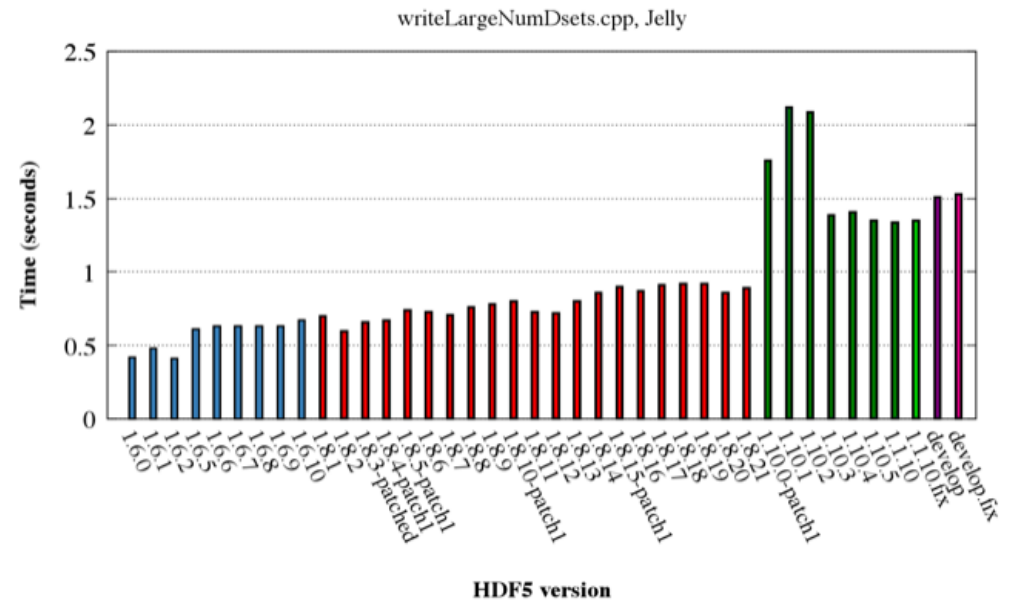
Writing 5K datasets with different versions of HDF5

HDF5 Roadmap for 2019-2020 and beyond

CGNS Application Before and After Tuning

Courtesy Greg Sjaardema, Sandia National Labs

How you can help us

The HDF Group

- Contact The HDF Group ASAP when you see performance degradation
- Submit application I/O kernel to our performance regression test suite
- Contact us
  - help@hdfgroup.org
  - prioritysupport@hdfgroup.org for Enterprise Support customers

HDF5 Development Directions

- Extending the current way of performing I/O
  - One process writing/reading to/from HDF5 file(s) (sequential)
  - Multiple processes writing/reading to/from shared file using MPI I/O (parallel)
- Extending HDF5 capabilities with
  - Virtual Object Layer (VOL)
  - Virtual File Drivers (VFD)
- Accessing HDF5 data on non-POSIX systems
  - Cloud
  - Object Store
  - Burst buffers
- More on the following slides

HDF5 Major Capabilities

HDF5 1.10 and 1.12. releases

Overview

- Parallel HDF5 enhancements
    - Scalability for file open/close
    - Collective metadata I/O
    - Writing compressed data in parallel
- Watch for a Webinar announcement later this Fall

Overview

- Sequential HDF5 enhancements
  - Driven and funded by Light Source Community
    ‣ Dynamically loaded compression filters
    ‣ Direct access to chunked data to speed up I/O (direct chunk write/read)
    ‣ Access to data stored in multiple files and datasets (Virtual Dataset or VDS)
    ‣ Access to HDF5 file under construction (Single Writer/Multiple Reader or SWMR)
  - File space management
    ‣ Reusing free space in the file
  - Paged allocation and page buffering for better I/O
  - Control over HDF5 "file version" (i.e., which version of the library can read this HDF5 file?)
  - ….

Where to find information about new releases?
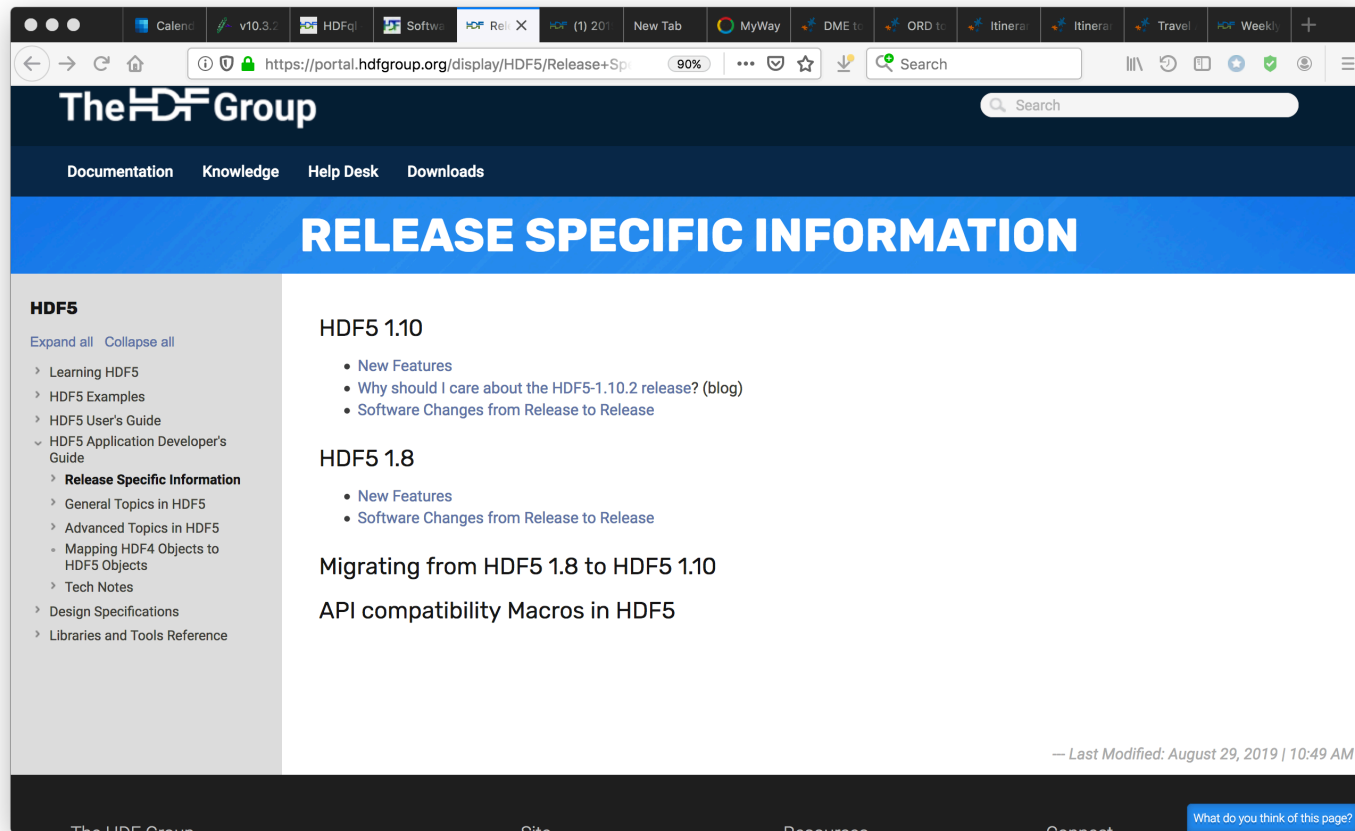
HDF5 Application Developer's Guide

Release Specific Information

Motivation for HDF5 Virtual Dataset

Dataset /VDS

```
 1  1  1  1  1  1
 2  2  2  2  2  2
 3  3  3  3  3  3
-1 -1 -1 -1 -1 -1
```

vds.h5

Dataset /B

```
2 2 2 2 2 2
```

b.h5

Dataset /A

```
1 1 1 1 1 1
```

a.h5

Dataset /C

```
3 3 3 3 3 3
```

c.h5

Data the application will see when reading VDS dataset from vds.h5
The last row is filled with the fill value.

Use Cases for Unlimited Datasets

HDF5 Roadmap for 2019-2020 and beyond
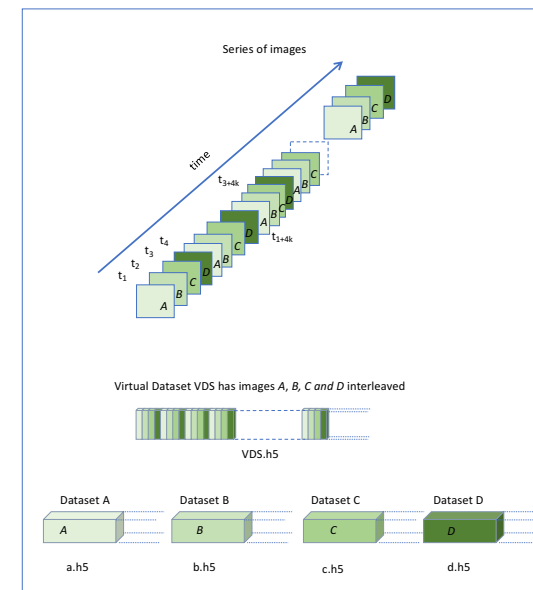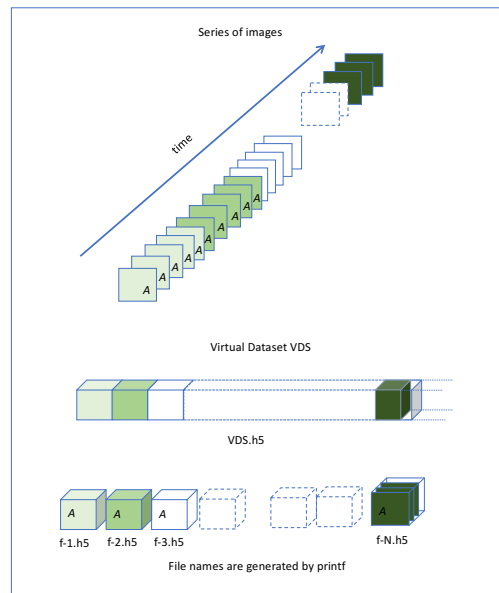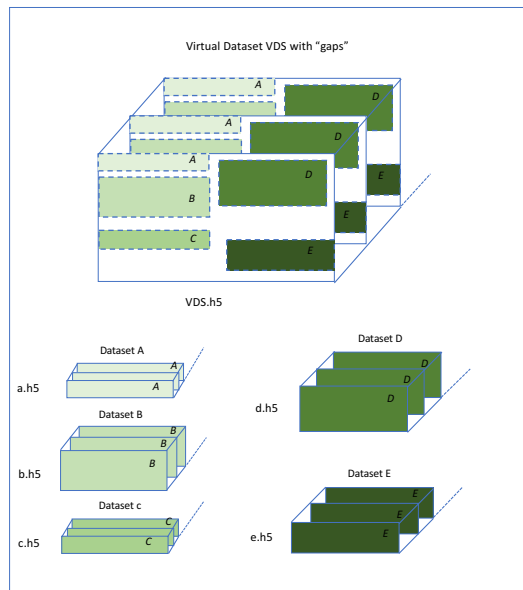
HDF5 Virtual Dataset

- New storage property introduce in HDF5 1.10.0
- Allows to "aggregate" data stored in multiple HDF5 datasets and files
- Works with SWMR
- Not a performant feature
  - Possible improvements
    ‣ Hyperslab selections performance
    ‣ Caching open files
    ‣ More mapping use cases
      ‣ Example: "snake" mapping

VS.

Source Datasets

File: a.h5
Dataset: /A
Dimensions: {15, 10, 10}

File: b.h5
Dataset: /B
Dimensions: {15, 10, 10}

File: c.h5
Dataset: /C
Dimensions: {10, 10, 15}

File: d.h5
Dataset: /D
Dimensions: {10, 15, 10}

Virtual Dataset

Dimensions: {10, 20, 20}
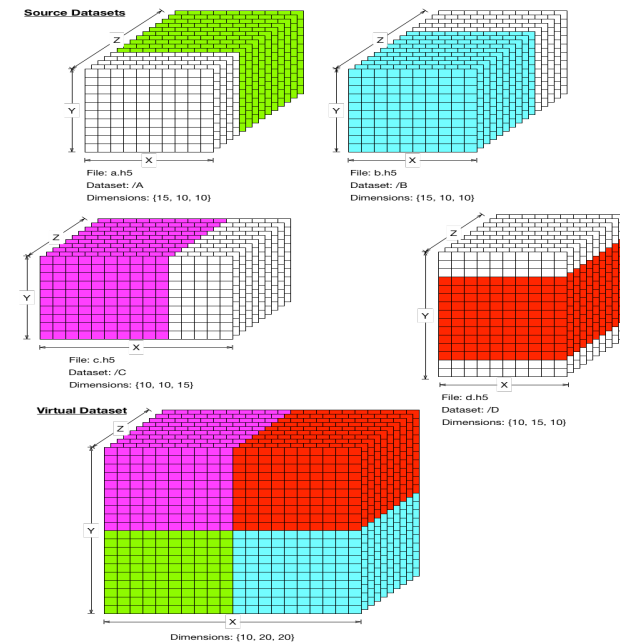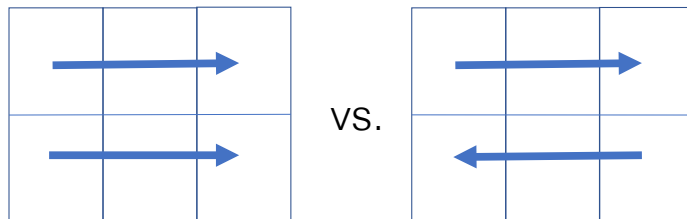
Single Writer/Multiple Reader (SWMR)

- Introduced in HDF5 1.10.0
- Implementation has many limitations
  - Based on POSIX semantics
    - No support for NSF
  - Works for raw data only
    - File structure cannot change
  - New data may not be visible for some time
  - Not high-performing
  - Doesn't work for parallel applications
  - Introduced file locking
  - Order of writer and reader is critical

New data
elements
…

Writer

Reader

…which can be read
by a reader…

with no IPC
necessary.

… are added
to a dataset
in the file…

HDF5 File
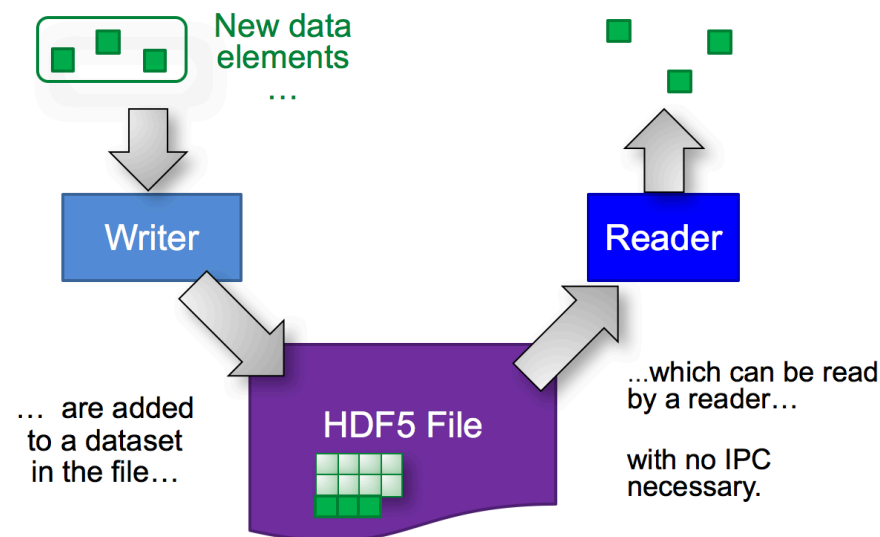
Full Single Writer/Multiple Reader (SWMR)

- **Goals**
  - Drop POSIX semantics requirement
    - ‣ Support for NFS
  - Allow modification of a file structure
    - ‣ Adding groups, links, datasets and attributes
    - ‣ Adding support for VL types
  - Guarantee maximum time from write availability to availability for read
  - Minimize maintenance cost
    - ‣ SWMR VFD

New data elements …

Writer

Reader

… are added to a dataset in the file…

HDF5 File

…which can be read by a reader…

with no IPC necessary.
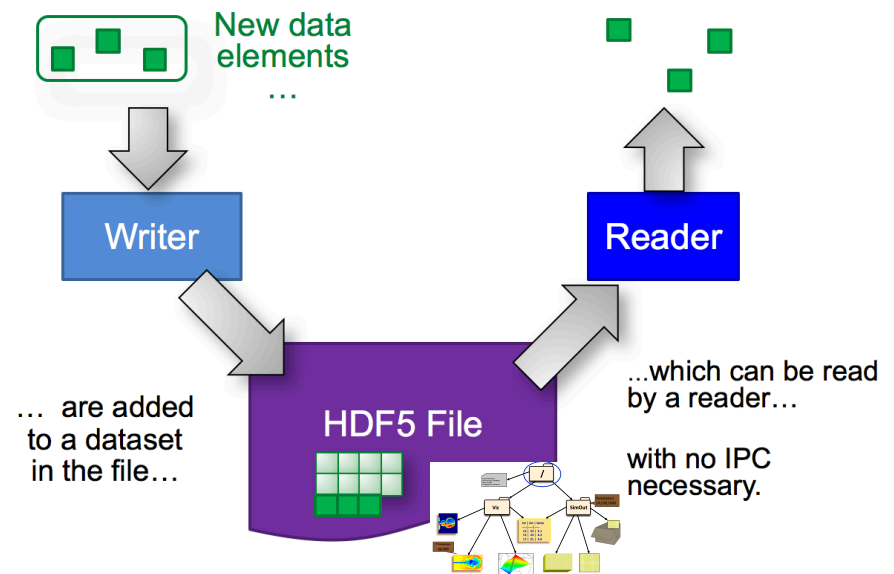
Demo

New SWMR implementation

Page buffering/ Paged allocation

- Introduced in 1.10.0

- Aggregate and align metadata and small data, and perform I/O on aligned pages

- See File Space Management Documentation https://portal.hdfgroup.org/display/HDF5/File+Space+Management

‣ One can reuse free space in the file between file open/close

Data and Metadata Aggregators

The aggregators pack small raw data and metadata allocations into aligned blocks (pages) which work with the page buffer

Data   Metadata

HDF5 File

Small allocations   Page size is tunable
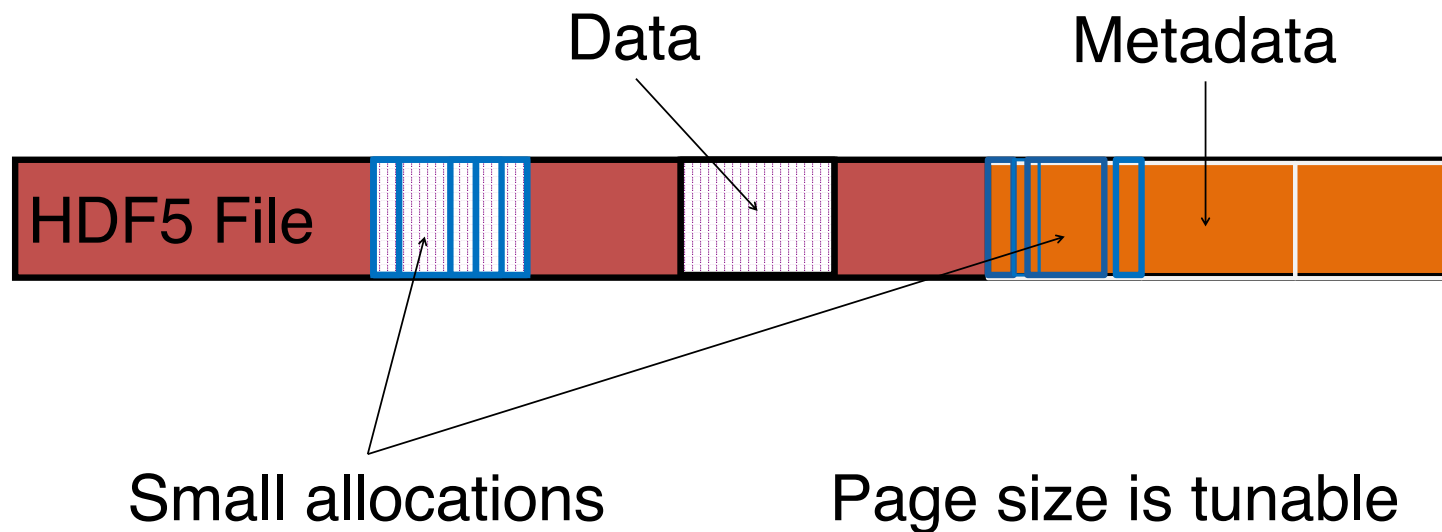
HDF5 Page Buffering

Page buffer contains MD and small data pages (L2 cache)

Pages are aligned

HDF5 File

HDF5 Page Buffering

HDF5 APIs

Internals, MDC cache, preparation for I/O

L2 cache

I/O is performed on pages

HDF5 File

HDF5 Library

VFD

HDF5 Roadmap for 2019-2020 and beyond
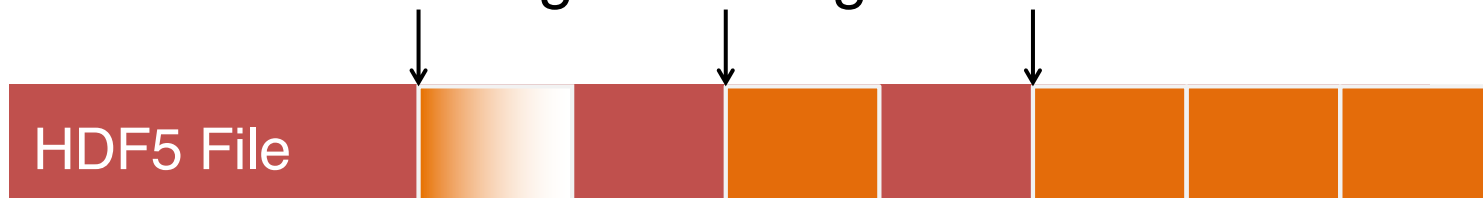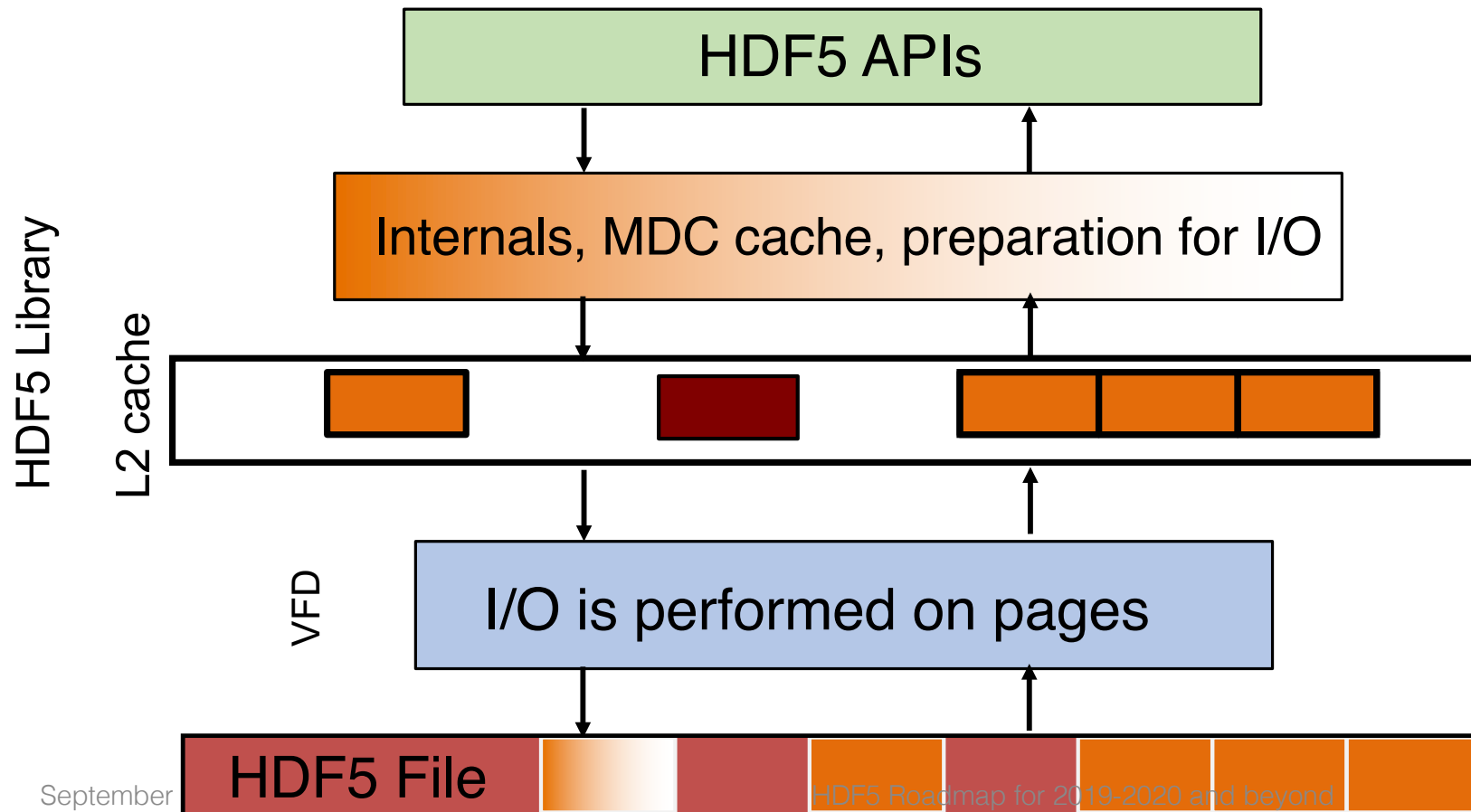
HDF5 Reference Datatype in 1.12.0

HDF5 dataset may store references to other objects or references to selected elements of a dataset

B.h5

Time step 36,000

```
V1 | V2  | temp
----|-----|-----
12 | 23 | 3.1
15 | 24 | 4.2
17 | 21 | 3.6
```

References to objects in other files are coming in HDF5 1.12.0

A.h5
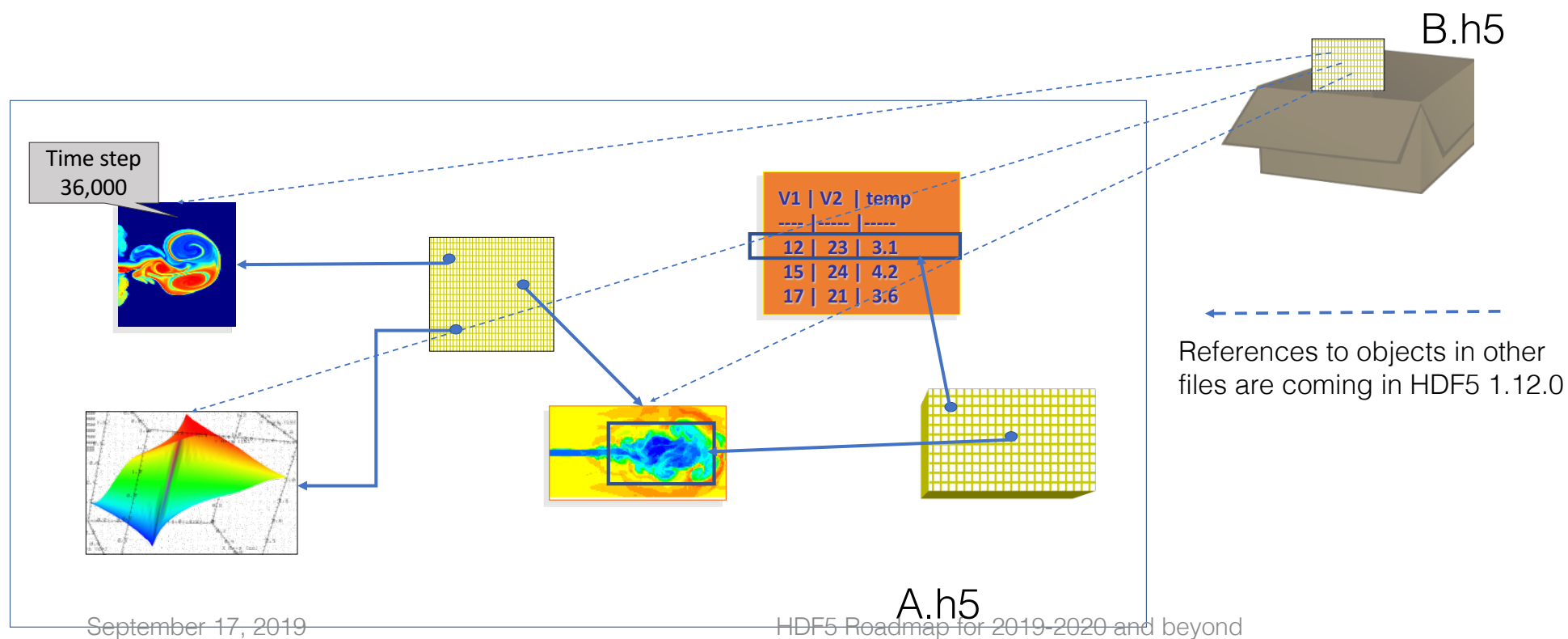
HDF5 Reference to Attribute Datatype

HDF5 1.12.0 will allow to reference an attribute of a dataset or a group

Time step 100,000

B.h5

Time step 36,000

V1	V2	temp
12	23	3.1
15	24	4.2
17	21	3.6

References to attributes are coming in HDF5 1.12.0

A.h5

The HDF Group
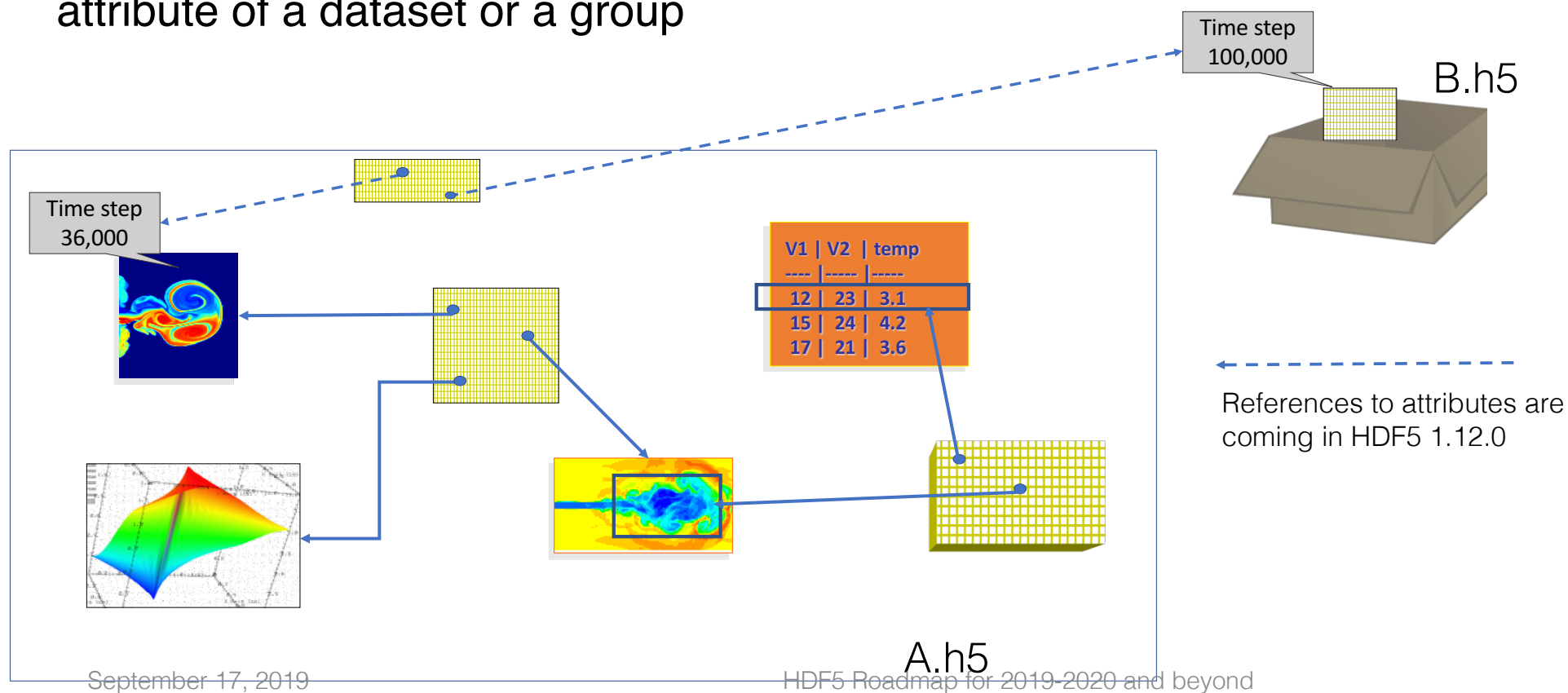
- **Motivation**

  - Our user and customers including vendors invested in HDF5 and HDF5-based software, but now HDF5 file format shows its limitations

  - Everyone likes HDF5 data model for its simplicity and versatility

  - Questions we asked ourselves:

    ‣ How current HDF5 applications can take advantage of new storage paradigms?

    ‣ How HDF5 data can be used with Big Data Technologies like Spark and Hadoop?

    ‣ How can we take advantage of new storage paradigms to improve I/O performance?

    ‣ How can we provide access to HDF5 data on new storage like Object Store and Cloud?

    ‣ *Can we extend current HDF5 to answer the above questions with minimum modifications to HDF5 applications?*

HDF5 1.12 Library Architecture

The HDF Group

HDF5 APIs
VOL Layer

VOL Connector

***

VOL Connector

VOL Connector

***

HDF5 Core

Virtual File Layer

Custom

VFD | VFD

***

VFD

Storage

Object Store   File system   Cloud
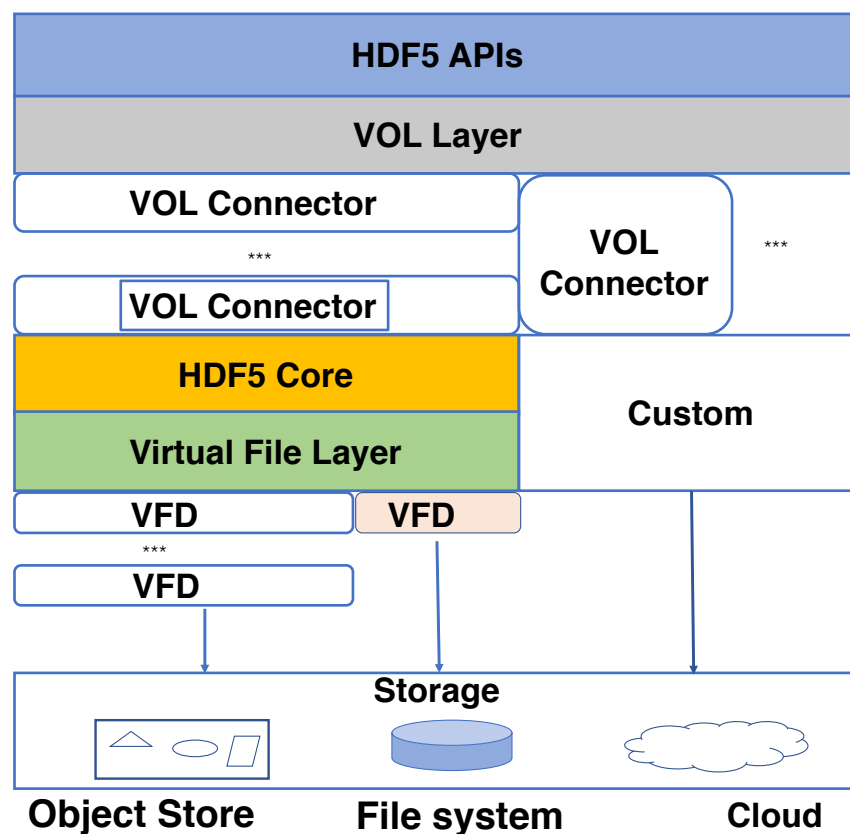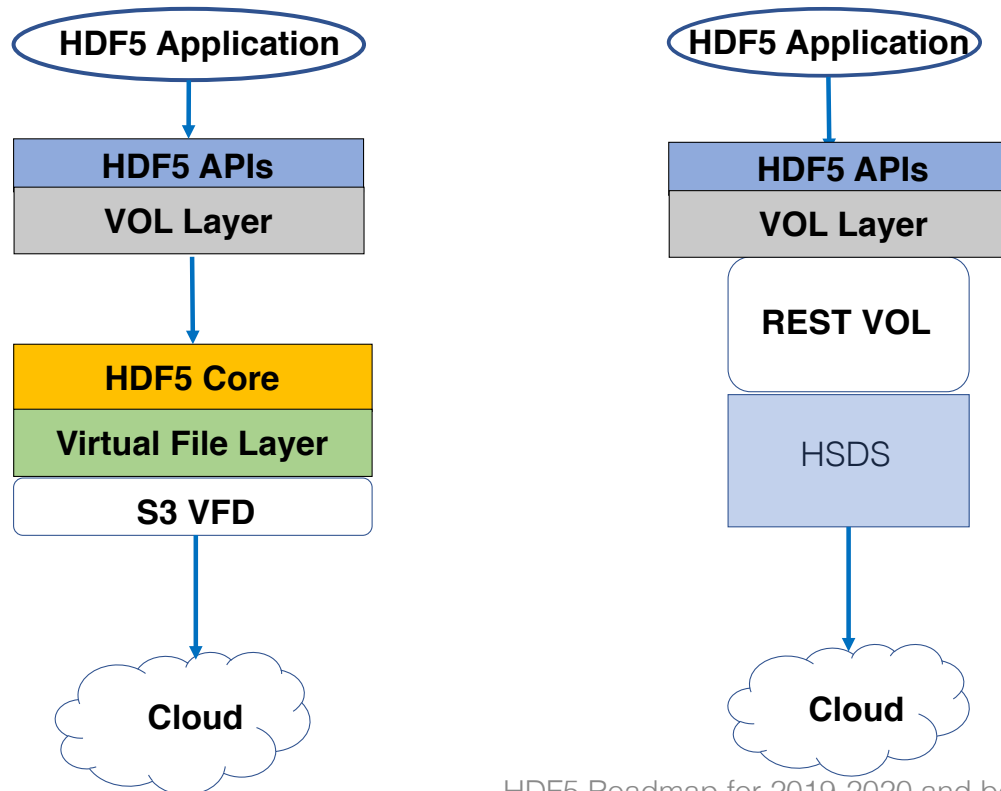
Colored shapes represent library components available in the current releases

HDF5 Applications Capabilities

Minor or no changes to applications are required to store data in the Cloud or Object Store

HDF5 Application

HDF5 APIs
VOL Layer

HDF5 Core
Virtual File Layer
S3 VFD

Cloud

HDF5 Application

HDF5 APIs
VOL Layer

REST VOL

HSDS

Cloud

HDF5 VOL Connectors

■ **https://bitbucket.hdfgroup.org/projects/HDF5VOL**

HDF5 Roadmap for 2019-2020 and beyond
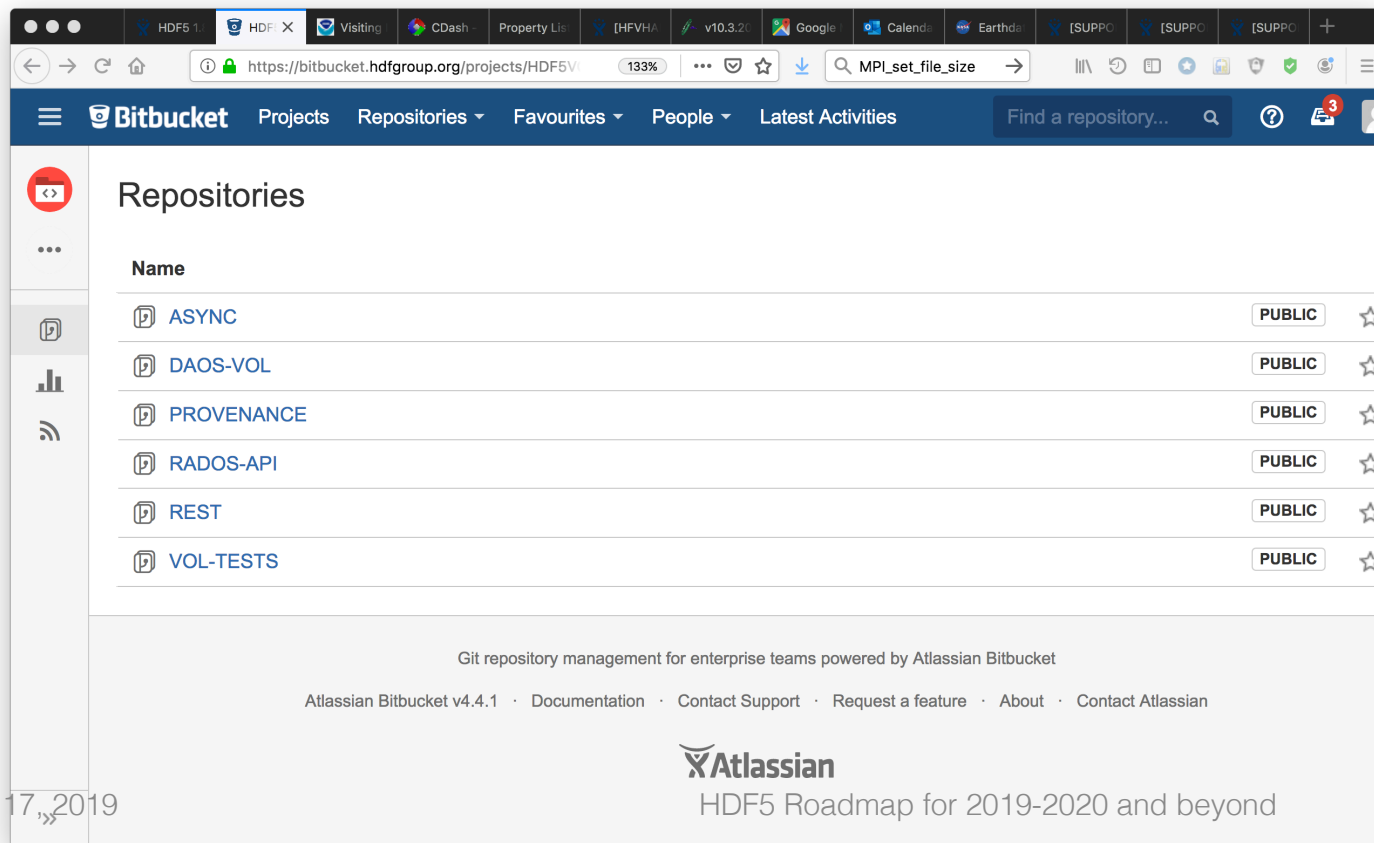
HDF5 VOL Connectors (cont'd)

The HDF Group

- Watch for a Webex on how to create a VOL connector

S3 VFD

The HDF Group

- HDF5 1.12.0 will have a VFD to access HDF5 file via Amazon Simple Storage Service (Amazon S3)
  - Requires minimum changes to the application code
- h5dump and h5ls tools have a flag to access HDF5 file on S3

  `h5ls --vfd=ros3`

  `https://s3.us-east-2.amazonaws.com/file.h5`

- Uses "range get" commands to get "bytes" from HDF5 file stored on S3
- New API to set up S3 VFD

  herr_t H5Pset_fapl_ros3(hid_t fapl_id, H5FD_ros3_fapl_t *fa)

  - Credentials are passed via parameter to the function

HDFS VFD

- HDF5 1.12.0 will have a VFD to access HDF5 file on Hadoop Distributed File System (HDFS)

- New API to access HDF5 file on HDFS

  herr_t H5Pset_fapl_hdfs(hid_t fapl_id);

- HDF5 command line tools with enabled HDFS VFD allows to extract metadata and raw data from HDF5 and netCDF4 files on HDFS, and use Hadoop streaming to collect data from multiple HDF5 files
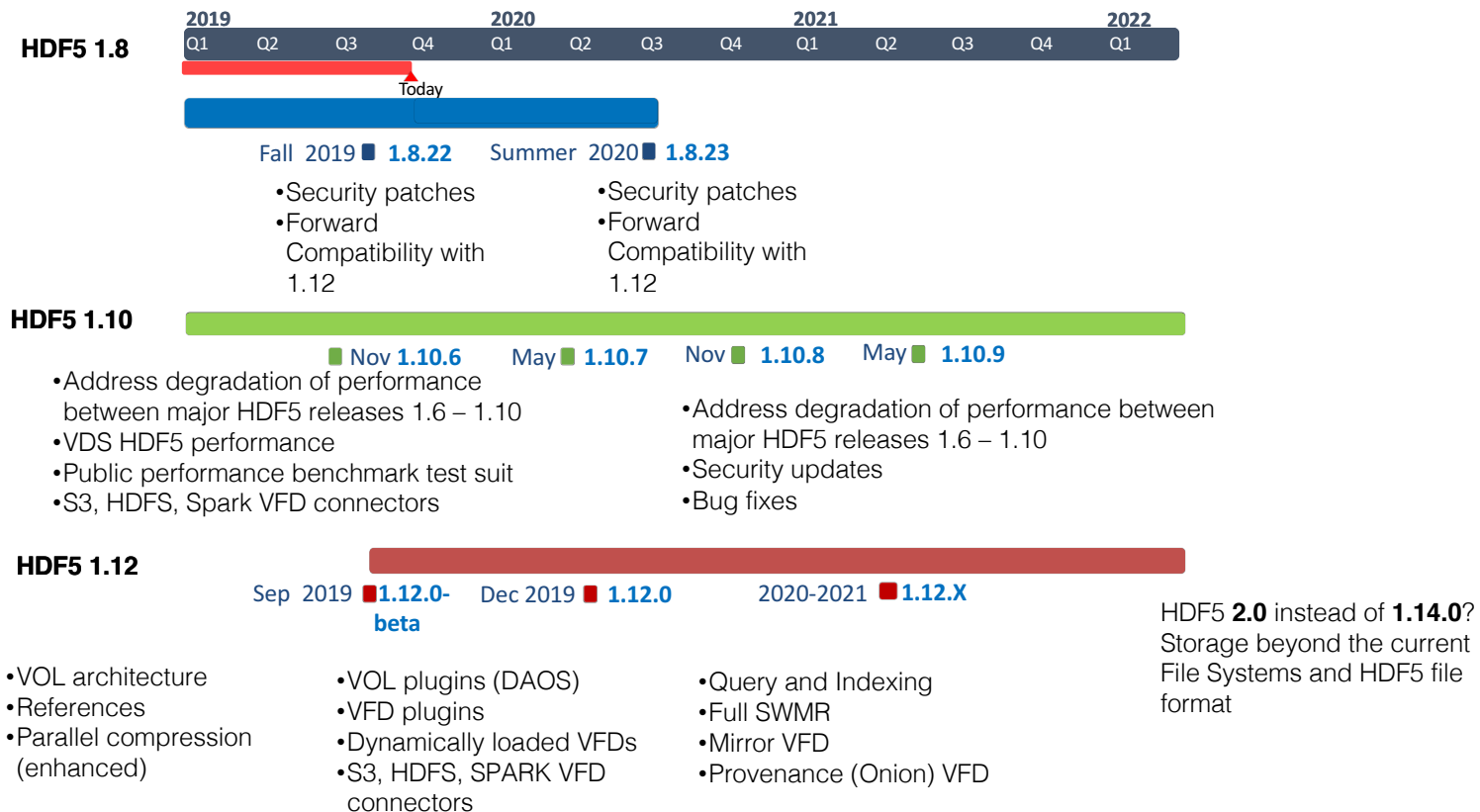
HDF5 2019-2023

Roadmap

HDF5 Roadmap 2019 - 2020

	2019				2020				2021				2022
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1

HDF5 1.8

Today

Fall 2019 ■ **1.8.22**   Summer 2020 ■ **1.8.23**

- •Security patches
- •Forward Compatibility with 1.12

- •Security patches
- •Forward Compatibility with 1.12

HDF5 1.10

■ Nov **1.10.6**   May ■ **1.10.7**   Nov ■ **1.10.8**   May ■ **1.10.9**

- •Address degradation of performance between major HDF5 releases 1.6 – 1.10
- •VDS HDF5 performance
- •Public performance benchmark test suit
- •S3, HDFS, Spark VFD connectors

- •Address degradation of performance between major HDF5 releases 1.6 – 1.10
- •Security updates
- •Bug fixes

HDF5 1.12

Sep 2019 ■**1.12.0-beta**   Dec 2019 ■ **1.12.0**   2020-2021 ■**1.12.X**

HDF5 **2.0** instead of **1.14.0**? Storage beyond the current File Systems and HDF5 file format

- •VOL architecture
- •References
- •Parallel compression (enhanced)

- •VOL plugins (DAOS)
- •VFD plugins
- •Dynamically loaded VFDs
- •S3, HDFS, SPARK VFD connectors

- •Query and Indexing
- •Full SWMR
- •Mirror VFD
- •Provenance (Onion) VFD
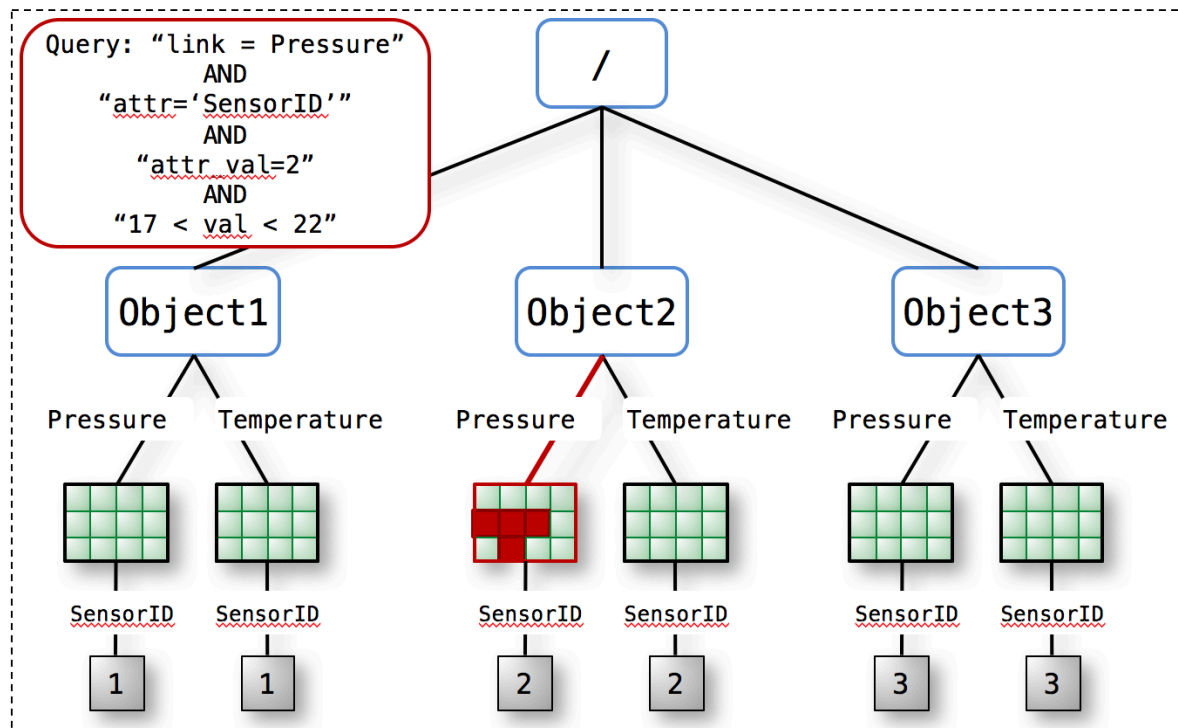
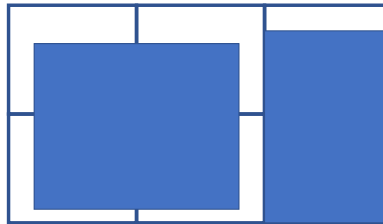# New features for Experimental and Observational Sciences

Mirror VFD
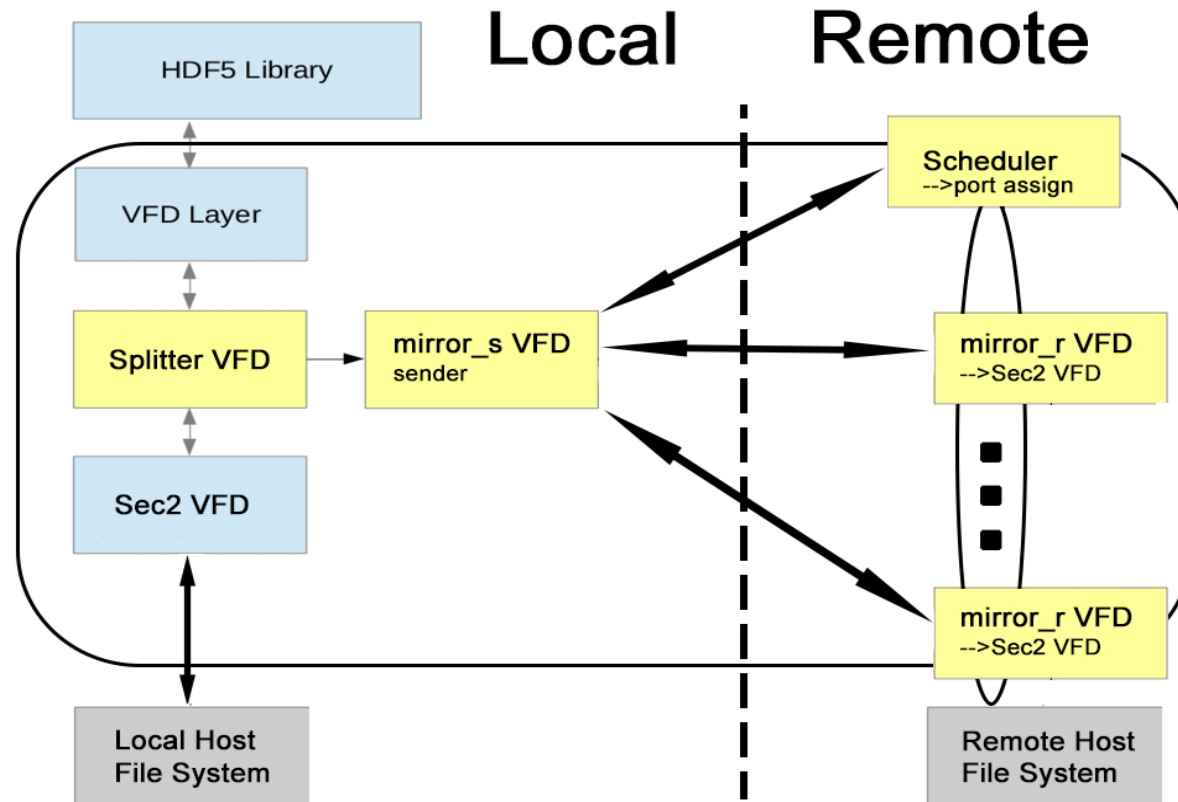
Query and Indexing Capabilities

HDF5 Query and Indexing will be release in HDF5 in 2020
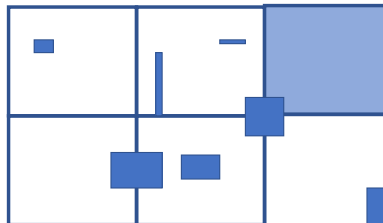
Sparse Storage

- **HDF5 was designed for "dense" array storage, when most of the array elements are defined:**

- **How to store sparse array efficiently? How to find the array elements that are defined?**
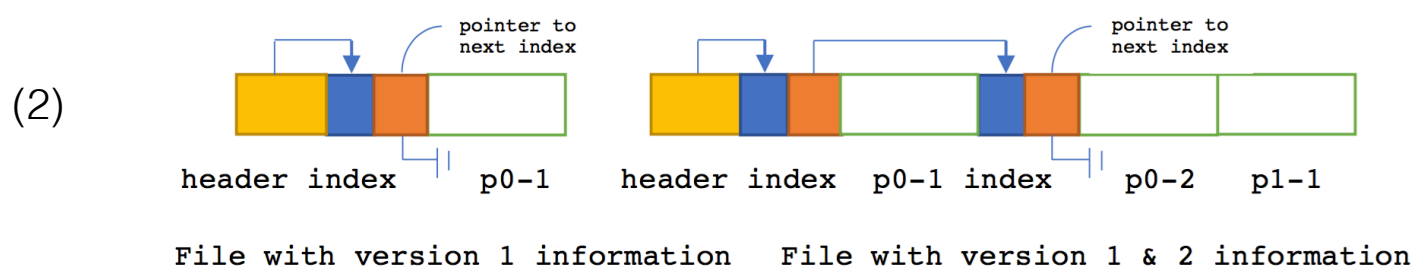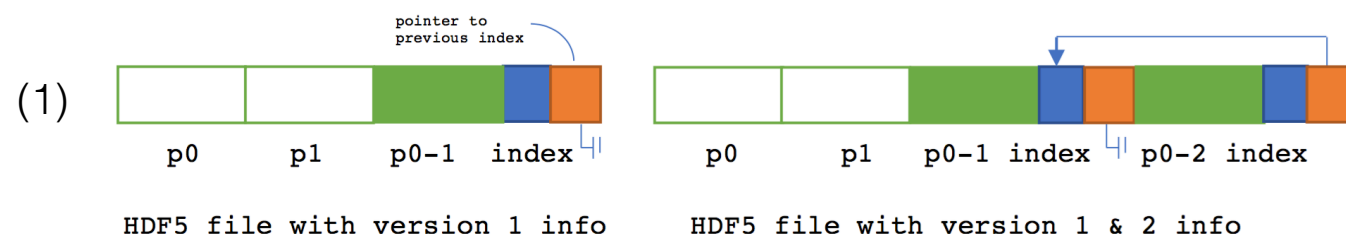

- **Share with us your use cases!**

HDF5 File Provenance

- HDF5 offers no support for version control and provenance management
- Data provenance is a major concern in cases where
  - Original data must be preserved
  - Changes are tracked and attributed
- General HDF5 "undo" capability is technically challenging
- Tracking changes per file open/close cycle may be sufficient
  - Divide HDF5 file linear byte space into pages
  - Build index that maps offsets in the files to the pages
  - Save modified pages in the new locations and have new index associated with the modified file
- There are several approaches to where we store modified pages and indexes.

HDF5 File Provenance

(1)

pointer to previous index

p0   p1   p0-1   index

HDF5 file with version 1 info

p0   p1   p0-1 index   p0-2 index

HDF5 file with version 1 & 2 info

(2)

pointer to next index

header index   p0-1

File with version 1 information

pointer to next index

header index   p0-1 index   p0-2   p1-1

File with version 1 & 2 information

HDF5 file paged allocation is used to track changed pages:
1. Tracking in the same file
2. Additional file with tracked pages
3. Additional files for each version

(3)

index   p0-1

Additional file for Version 1

index p0-2   p1-1

Additional file for Version 2
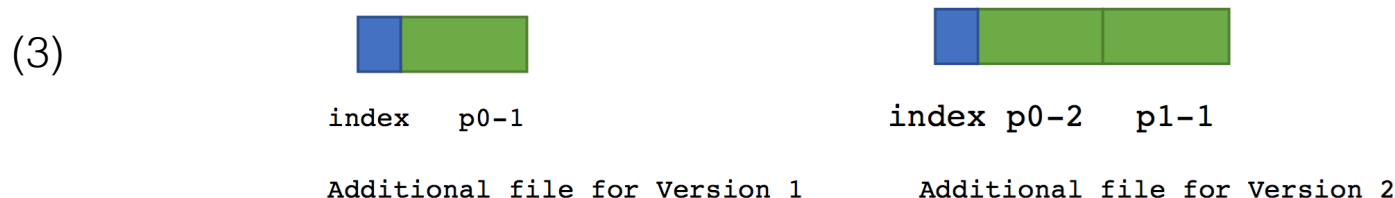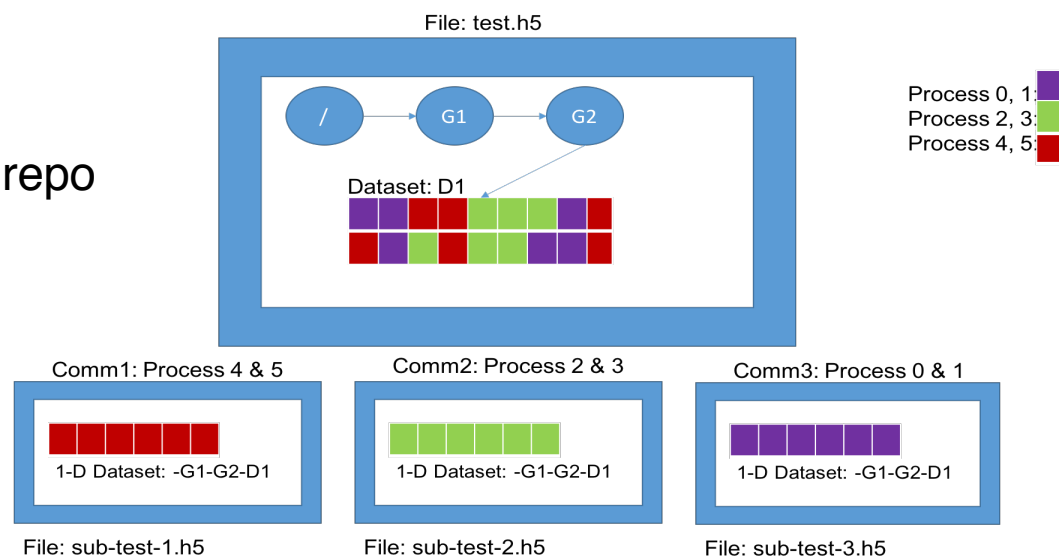
New features for HDF5 Parallel Lbrary

Sub-filing

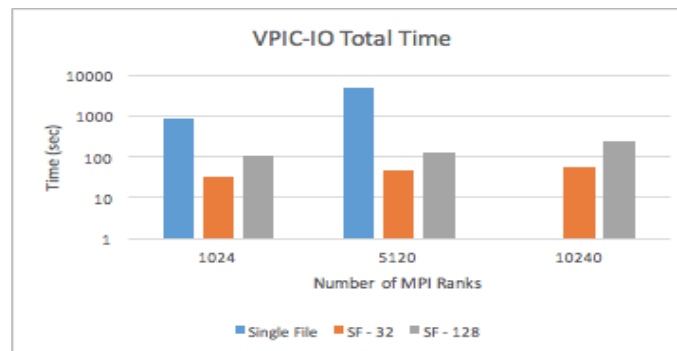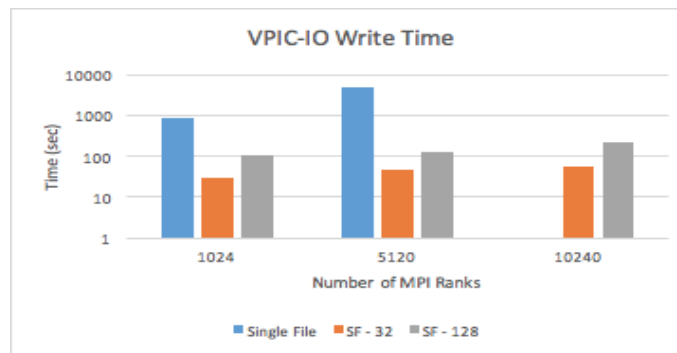HDF5 Roadmap for 2019-2020 and beyond

HDF5 Sub-filing

The HDF Group

- Sub-filing
  - M processes write to N files to reduce lock and contention issues on PFS
  - Compromise between shared file and file/per process approach
- VDS was used for prototyping HDF5 sub-filing feature
  - Transparent to application
  - Promising performance
  - See "subfiling" branch in HDF5 repo

File: test.h5

/ → G1 → G2

Process 0, 1:
Process 2, 3:
Process 4, 5:

Dataset: D1

Comm1: Process 4 & 5

1-D Dataset: -G1-G2-D1

File: sub-test-1.h5

Comm2: Process 2 & 3

1-D Dataset: -G1-G2-D1

File: sub-test-2.h5

Comm3: Process 0 & 1

1-D Dataset: -G1-G2-D1

File: sub-test-3.h5

HDF5 Sub-filing (cont'd)

- Initial performance results are promising

Thank you!

Questions?