HDF5 Future Work Ideas
HUG23

August 16-18, 2023

The HDF Group

Dana Robinson
Director of Software Engineering
The HDF Group

Big Things

What "big picture" things are next?

The HDF Group

The last several years of HDF5 development have strongly focused on HPC needs due to the Exascale Computing Project

- The Virtual Object Layer and associated HPC connectors
- Parallel I/O improvements
- HPC testing and bug fixing

Where do we go from here?

What are the next big challenges for HDF5 that we should be working on?

Journaling / Crash Integrity

The HDF Group

- File corruption on writer crashes is still a problem

Unicode

- Needs a global survey of where and how Unicode should work in the library
- Biggest problem is Windows, where UTF-8 isn't the native way to do Unicode

Multithreading Support

The HDF Group

- Being worked on by Lifeboat

Streaming

The HDF Group

- What's the most efficient way to get streaming data into HDF5?
- Do we need different file data structures?

Sparse Data

- Being worked on by Lifeboat

Better support for variable-length data

The HDF Group

- Stop storing data in metadata structures (global heaps)
- Find a way to make common use cases fast
- Enable compression

Cloud storage

The HDF Group

- Create cloud-optimized HDF5 files by default
- Document cloud best practices in the HDF5 User's Guide
- REST VOL becomes a first-class citizen
- ros3 VFD improvements
  - Performance
  - Expand to Google, Azure
  - Move away from cURL
  - Use selection I/O
- Improve h5repack performance

Useful Error Reporting

The HDF Group

- Current scheme is not great
  - herr_t only has two values
  - Error stacks are inconsistent and difficult to parse
- Expanding herr_t values to include values for "things a user can do something about" would be nice
  - Out of memory
  - File access problems
  - Invalid parameters
- Could report our usual -1 value as "library internal error"
- It'd be nice to simplify the internal error macros, too (lots of repetition)

Smaller Things

Smaller Things

The HDF Group

Aside from the "big picture" things mentioned on the earlier slides, there are a lot of smaller-scale things that should be done in the library

Improved Performance

The HDF Group

- We need a performance regression test harness
  - Need to identify critical workflows
  - Make them faster
  - Regression tests to ensure they stay fast
- Should create a performance section in the HDF5 User's Guide
  - Detail best practices and anti-patterns
- Think about changing library defaults to work better on modern hardware
  - Default cache sizes, etc.
  - Default library file format version

Expand Data Types

- Boolean
- Float16 (and other edge-AI-oriented types?)

Improved Security

The HDF Group

- Continued improvement of the HDF5 CVE test suite
  - Create proof-of-vulnerability files for missing Talos CVEs
  - Test GIF tool vulnerabilities
  - Add to HDF5 repo as an action
- Fix GIF tool vulnerabilities
  - Currently bypassed by not building them
  - Might also move to another repo

Better Code Structure

The HDF Group

- Better encapsulation
- Break down large packages into smaller units
- Reduce dependencies
  - Especially "friends"
- Turn HDFG-isms into normal C code
  - HD prefixed C/POSIX calls
  - hbool_t

More Automation

The HDF Group

- More automated testing
  - VOL connectors
  - Binary compatibility
  - VFD checks
- Project management, releases
  - Set up potential merges

More Transparency

The HDF Group

- Move remaining Jira issues to GitHub (in progress)
- Fill out GitHub wiki section
- Product-level project management done via GitHub

Support Our Non-Profit Mission

The HDF Group

To ensure efficient and equitable access to science and engineering data across platforms and environments, now and forever.

THESE DON'T COST YOU A DIME
Help Desk Support
Sustaining Engineering
HDF Clinic, Working Group
Webinars, User Events
HDF User Forum
Community Outreach
Assured Longevity of HDF Technologies

HELP US TO KEEP IT THAT WAY
Become a Code Owner
Consult with Us
Purchase Custom Development
Get HDF Software Priority Support
Donate or be a Sponsor
Collaborate with Us on a Proposal
Become an HDF Advocate

Contact: info@hdfgroup.org

https://www.hdfgroup.org/donate

THANK YOU!

Questions & Comments?