Through the Looking Glass: creating an HDF data prism

Matthew Dougherty
Consortium for Management of Experimental Data in Structural Biology
May 22, 2008

Hierarchical Data Format

- Developed by NCSA
- Maintained & developed by the HDFgroup
- Supported as BSD open source
- Scientific data format
- "File system within a file system"
- Management of large heterogeneous datasets
- Parallel IO
- Archival
- ISO/STEP

Filesystem in **USE**r space

With FUSE it is possible to implement a fully functional filesystem in a userspace program. Features include:

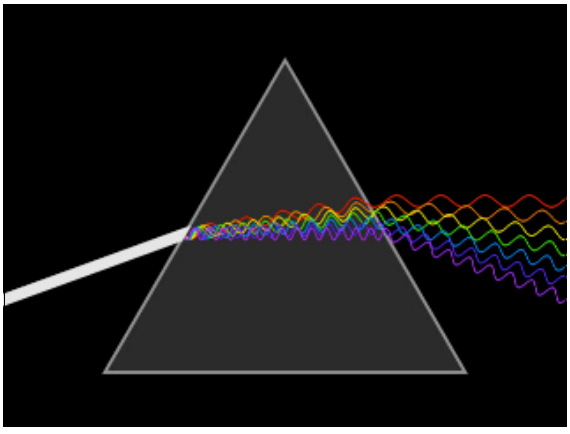
- Simple installation (no need to patch or recompile the kernel)
- Secure implementation
- Userspace - kernel interface is very efficient
- Usable by non privileged users
- Runs on Linux, Mac, FreeBSD, MS Windows, Solaris, GNU, NetBSD
- Has proven very stable over time
- Simple library API, implementing a filesystem is simple, a hello world filesystem is less than a 100 lines long.
- http://fuse.sourceforge.net

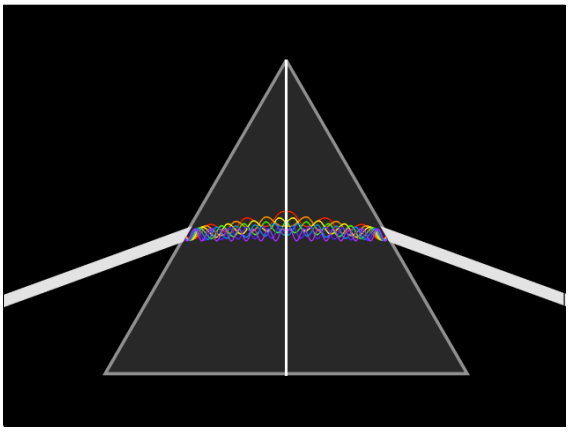
First HDF-FUSE prototype

- MacFUSE, Amit Singh/Google
- Ability to click on an HDF file to mount it
- Ability to read datasets in group /FUSE
- One user file becomes one HDF dataset
- Datasets were 1D byte
- Regular applications are unable to detect any filesystem difference
- Ability to amalgamate non-image datasets into HDF,
- Intended files: PDF, word, excel, project

What are the implications?

- The mirror and the prism paradigm





What are the implications?

- The mirror and the prism paradigm
- Integrated parallel file system
- Transparent integration of HDF into any application.
- 1D byte datasets could become optimal nD image datasets, if the image format was known during intake, allowing for conversion and metadata segmentation.
- Namespace management
- A project management pipeline could be enforced in this parallel file system.

What are the implications?

- Elimination of pixel redundancies, mass storage savings
- Transparent integration of Compression
- Transparent integration of Hyperslabs
- Transparent integration of Chunking
- New strategies and paradigm shifts could be developed.
  - DICOM
  - nD Scientific Image Format (nD-SIF)
  - Scientific Acquisition, Visualization, Repository Environment (SAVRE)

nom de guerre ImageCore

- **ImageCore** is a proposed methodology to organize images and arbitrary metadata within an HDF file. It can be used as a generic image format, or it can be linked into existing image data formats that implement HDF.
- Through the use of FUSE, application programs can transparently participate, removing a major obstacle of adoption.

ImageCore

1) Define a reserved HDF namespace for image-pixels using a simple naming convention.
2) Associate images, pixels and metadata using the Resource Description Framework (RDF).
3) Provide an optional reserved IC namespace for grouping adjunct user or application metadata.
4) Provide integration protocols for images and metadata that exists in user or application namespaces.

ImageCore

1) An HDF root group "/imageCore/", forming the primary IC namespace.
2) A *chronological* RDF image log, having the reserved HDF dataset name "/imageCore/0". It is an nx4 extensible table of time-stamped RDF values
3) An expandable number of N-dimensional images whose pixels are contained in HDF datasets.
4) Optional metadata can be grouped in the IC namespace corresponding to the image datasets names. These adjunct metadata may contain any type of user datasets, groups, and attributes

Avenues of adoption

- MEDSBIO / IUCr
- TDWG / IUBS
- NISO

Organizational support

- NSF
- NIH/NCRR
- NIST
- DOE
- DARPA
- LOC/NARA
- Private foundations
- elixir-europe.org
- instruct-fp7.eu