HDF-VFS

Matthew Dougherty
National Center for Macromolecular Imaging
Baylor College of Medicine, Houston, Texas

MOTIVATION

- advanced 3D spatial EM imagery (1998)
- pedestrian MPEG video imagery (2004)
- all project and animation intermediate files (2008)

REQUIREMENTS

- HPC
- Interoperability
- Preservation
- Legacy transparency

MECHANISMS

- HDF-VFS
- Performance
- Provenance
- Community extensions
- Automated project management
- Registries (CPID, communities, s/w)

ESTABLISHED COMMUNITIES

- EOS
- NeXus
- Matlab
- IDL
- netCDF

EVOLVING COMMUNITIES

- EM
- Bio-VIZ
- X-ray crystallography
- Optical microscopy
- Astrophysics
- Data storage
- Astronomy
- Genomics

OUTREACH

- ACM
- NIST
- NSF PROPOSALS
- IWGDD

TRIGGER

- Animation production
- EM databank/PDB/EBI/NSF

The PLAN

- Phase one

  <HDF-VFS attribute>

  -

- Phase two

  /groups/

  {datasets}

  - Performance
  - Provenance
  - Community filters

<hdf-vfs>

- Version ID
- Visible flag
- Initial UUID
- Current UUID
- Image UUID
- Created timestamp
- Modified timestamp
- Accessed timestamp
- VFS file/directory name
- Concatenation link
- Permissions (read, permanent read, write, non-extensible write)
- Group to dataset inheritance
- Dataset size
- Provenance
- Performance
- Filter

Sleeping Gorilla issues

- Design, development, maintenance, long-term support
- Location and control of Code
  (sourceforge, google, NCMI, HDFgroup)
- HDFGroup overlap
- HDF file inside an HDF file
- Imagery RDF design
- Provenance RDF design
- Performance RDF design
- Filter & community registries
- Microsoft

Suggestions, comments

Matthew Dougherty

matthewd@bcm.edu