

HDF Update

Mike Folk

National Center for Supercomputing Applications

Science Data Processing Workshop

February 26-28, 2002

Topics

- What is HDF?
- HDF4 update and plans
- HDF5 update and plans

NCSA HDF Mission

To develop, promote, deploy, and support open and free technologies that facilitate scientific data storage, exchange, access, analysis and discovery.

What is HDF?

- Format and software for scientific data
- Stores images, arrays, tables, etc.
- Storage and I/O efficiency
- Free and commercial software support
- Promote standards
- Broad user base in engineering & science

Who is supporting HDF?

- NASA/ESDIS
 - Earth science applications, instrument data
 - All aspects of data management
- DOE/ASCI (Accelerated Strategic Computing Init.)
 - Simulations on massively parallel machines
 - Emphasis on parallel I/O performance, functionality
- DOE Scientific Data Analysis & Computation Program
 - High performance I/O R & D
- NCSA
 - Grid, Vis, other R&D, user support
- Others
 - Applications, support, some R&D

HDF4

Goals for HDF4

- Support Terra, Aqua & other EOS data users
- Maintain and upgrade library and tools
- Address differences between HDF4 & HDF5

HDF4 milestones since Sept 2000

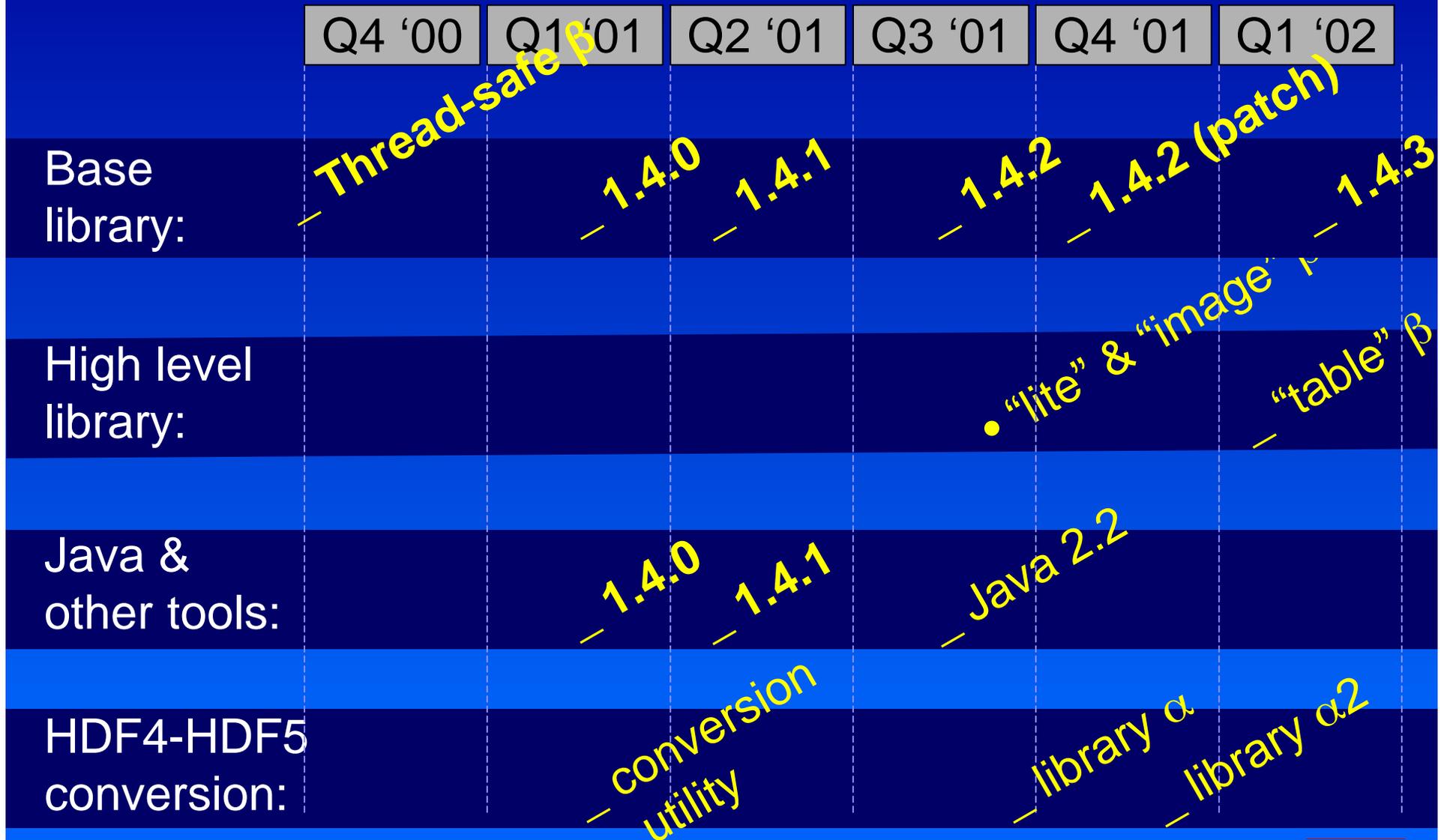
| | Q4 '00 | Q1 '01 | Q2 '01 | Q3 '01 | Q4 '01 | Q1 '02 |
|-----------------------------|--------|---------------|--------|--------|-------------|---------------|
| HDF4 library and utilities: | | • HDF4.1 r4 | | | • HDF4.1 r5 | |
| HDF4 Java API & tools: | | • Version 2.6 | | | | • Version 2.7 |

HDF4 releases

- HDF4.1 r4 (Nov 2000)
 - Chunking and chunking with compression
 - JPEG compression fixed
 - hdp utility enhancements
 - HDF4-to-GIF and GIF-to-HDF4 converters
- HDF4.1 r5 (Nov 2001)
 - Compression query functions
 - Vdata: set size of Vdata link blocks
 - Tools updates
 - Added Solaris 2.8
 - Retired Solaris 2.6, OSF1 V4.0 and HPUX 10.20

HDF5

HDF5 milestones since Sept 2000



HDF5 Library Work

New platforms and languages

- New platforms
 - HP-UX
 - IBM SP
 - IA64 architecture
 - Solaris 2.8 with 64-bit
 - Linux in parallel mode
 - Metrowerks Code Warrior Windows compiler
- Fortran 90
 - Solaris 2.6 & 2.7, O2K, DEC OSF, Linux, Cray T3E, SV1
 - Parallel on T3E & O2K
 - Windows
- C++
 - Solaris 2.6 and 2.7, Free BSD, Linux, Windows

New to HDF5

- Virtual File Layer – alternate I/O drivers
- Array datatype – array as atomic type
- File sizes greater than 2GB on Linux
- Performance improvements – parallel & serial
- Improvements in configuration
- Tutorials and documentation

Next major release -- HDF5 1.6

- New format and library features
 - Reclamation of free space within a file
 - Enhanced hyperslab/region selection
 - Dimension scale support
 - Bzip2 compression
 - Generic Properties
 - Performance improvements
- Parallel I/O performance benchmark suite
- Release date: Fall 2002

High level APIs

- HDF5 routines that do more operations per call than the basic HDF5 interface
- Goals
 - Make HDF5 easier to use
 - Encourage standard ways to store objects in HDF5

High level APIs

- Lite – done
- Image – done
- Table – partly done
- Dimension scale – in the works
- Unstructured grids – in the works
- http://hdf.ncsa.uiuc.edu/HDF5/hdf5_hl/doc/

HDF5 High Level APIs – **HDF5 Image**

- For datasets to be interpreted as images/palettes
 - 2-D raster data like HDF4 raster images
- Image operations
 - Create, write, read, query
- Based on “HDF5 Image & Palette Specification”

HDF5 High Level APIs – HDF5 Table

- For datasets to be interpreted as “tables”
 - A collection of records
 - All records have the same structure
 - Like Vdatas in HDF4, but more operations
- Table operations
 - Create, write, read, query
 - Insert, delete records or fields
 - Future: sort and search

HDF5 tools activities

H5View – new features

- Improved editing
- XML support
- Supports “image” data sets
- Line plots of row/column data
- Save data to a text file
- Set user preferences
- <http://hdf.ncsa.uiuc.edu/java-hdf5-html/>

Future Viewer work

- Objectives
 - Support transition from HDF4 to HDF5
 - Eliminate duplicate work on HDF4 and HDF5 tools
 - Provide plug-in architecture for tool development
- Three stages
 - Object model to cover both HDF4 & HDF5
 - Nearly done
 - Simple browser for both HDF4 and HDF5
 - Summer 2002
 - Editor
 - Fall 2002
 - Modular toolkit for browser plug-ins
 - Fall 2002, if all goes well

XML and web experiments

- Explore XML, Web applications for HDF5
 - Investigate use of XML
 - investigate Web XML technologies, use with HDF5
- XML DTD for HDF5
- Conversions involving XML
 - HDF4 → HDF5 → XML
 - Compression and timing study
 - netcdf to HDF5 translation, via XML style sheet
- XML Schema for HDF5 investigated

Other Java-inspired explorations

- Tomcat web server and JSP servlets
 - HDF5 file → XML → html → display in a web browser
 - Shows promise for providing remote access to HDF5
- HDF5 and CORBA
 - Investigating using CORBA with Java and HDF5.
 - Data access done in a CORBA servant written in C/C++
 - Browsing and presenting the information done in Java
 - Uses CORBA is alternative to the Java Native Interface, which calls C directly from Java.

Other Tools Activities

- HDF5-to-GIF and GIF-to-HDF5 converters
- H5dump subsetting

Facilitating the transition from HDF4 to HDF5

The transition from HDF4 to HDF5

- We support both HDF4 and HDF5.
- We will continue to maintain HDF4, as long as we are funded to do so.
- We recommend:
 - Use HDF5 for new projects
 - Consider migrating from HDF4 to HDF5 to take advantage of the improved features and performance of HDF5

HDF4-to-HDF5 Work

- Mapping specification
- Utility
- Library
- Experiments

HDF4-to-HDF5 Work

- Later talk by Bob McGrath to cover
 - The key technical challenges
 - NCSA toolkit work
 - Experiments to support transition

Other Activities of Interest

Parallel HDF5

- HDF5 supports parallel I/O
- Emphasis on performance
- Driven by DOE labs
- Tutorial

<http://hdf.ncsa.uiuc.edu/HDF5/doc/Tutor/>

- See Elena's talk

Parallel I/O benchmark suite

- A set of parallel performance benchmarks that can be distributed with HDF5
- Parallel HDF5 in MPI environment
- Measures performance of raw I/O, MPI-I/O, & HDF5
- Draft documentation:
http://hdf/RFC/PIO_Perf/PHDF5_performance.html

Other performance studies

See Elena's talk

Thread-Safe HDF5 (Beta)

- Uses Pthreads (POSIX threads) library.
- Implements Phase 1 of HDF5 thread safety plan
 - One lock per program
- Later phases
 - 2: Separate locks per major operation (read, write, convert)
 - 3: Separate locks for all operations.
- For details see
 - <http://hdf.ncsa.uiuc.edu/HDF5/papers/mthdf/>

HDF5-DODS* Server

- HDF5 data available from DODS clients
- Demonstrated with Ferret client
- Documentation
 - *HDF5-DODS Data Model and Mapping*
 - *The HDF5-DODS Server Prototype*
 - *Demo of HDF5-DODS Server Prototype with the DODS-Ferret Client*
- <http://hdf.ncsa.uiuc.edu/apps/dods/>

* Distributed Oceanographic Data System

Transform architecture prototype

- Provides a mechanism for operating on data as it is moved from one location to another
- Inspired by HDF5 I/O operations when moving data between memory and file:
 - Convert numbers – size, endianness, etc.
 - Selection – subsetting & subsampling
 - Linear order – row major vs. column major
- Later
 - Change value/Meaning of data – units, coordinate system, etc.

Outreach

- Documentation
 - Expanded and improved
 - Searchable, printable
 - HDF5 User's Guide
- Advanced and parallel HDF5 Tutorial
 - <http://hdf.ncsa.uiuc.edu/HDF5/doc/Tutor/>
 - Presented at SC2001
 - http://hdf.ncsa.uiuc.edu/HDF5/papers/SC2001/SC01_tutorial/

szip Compression

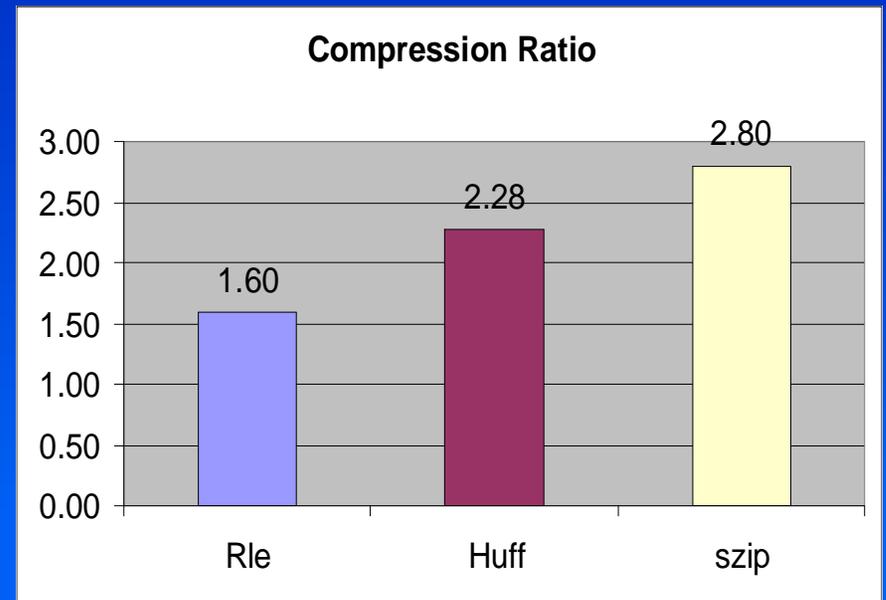
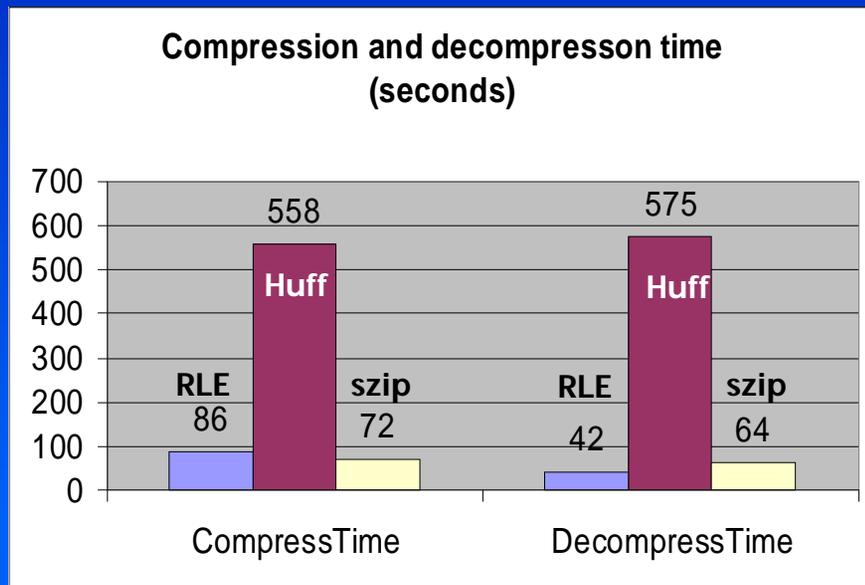
Reporting for
Pen-Shu Yeh and Wei Xia

Szip Compression in HDF 4

- CCSDS lossless compression
 - Fast, effective compression method for EOS data
 - Outperforms other common techniques in size & speed
- Szip software
 - Implements CCSDS lossless compression
 - Originally developed at University of New Mexico
 - Implemented in HDF-4 by UNM
- HDF-4 with CCSDS-szip option (HDF-4-Szip)
 - Delivered to GSFC and tested
 - Tested with MODIS Level-1B data from GSFC DAAC

Sample results

Convert 343 Megabyte MODIS file
Conversion performed on PC-Linux 7.1, Pentium II, 300Mhz



Szip-HDF4 software

- HDF4 library routine **SDsetcompress**
 - compress existing dataset or create new compressed dataset
- Utilities
 - **bin2hdf** – input array from binary file, compress using HDF-4-Szip, and output compressed data in HDF file
 - **hdf2chdf** – input data from HDF file, compress using HDF-4-Szip, and output compressed data in new HDF file
 - **chdf2bin** – input HDF file (with or without compression), and output user-selected arrays as separated binary files

Future Work

- Compress IEEE 32 bit float data with HDF-4-Szip
- Use HDF chunking routines
 - Especially beneficial for subsetting performance
- Test HDF-4-szip on different computer platforms
 - Currently HDF-4-Szip has been tested under Linux only
 - Others to test: Sun, SGI and HP, Windows 98 and NT
- Advocacy in the scientific community
 - Add User's Guides and documentation to HDF docs.
 - Present and demo at workshops and conferences.
- Resolve distribution issues with NCSA and UNM

Information Sources



- HDF website
 - <http://hdf.ncsa.uiuc.edu/>
- HDF5 Information Center
 - <http://hdf.ncsa.uiuc.edu/HDF5/>
- HDF Helpdesk
 - hdfhelp@ncsa.uiuc.edu
- HDF users mailing list
 - hdfnews@ncsa.uiuc.edu

