NetCDF-4: Software Implementing an Enhanced Data Model for the Geosciences

Russ Rew, Ed Hartnett, and John Caron

UCAR Unidata Program, Boulder

2006-01-31

Acknowledgments

- This work was supported by the NASA Earth Science Technology Office under NASA award AIST-02-0071.
- Unidata's work is primarily supported by the National Science Foundation.
- We appreciate the collaboration and development efforts of the NCSA HDF Group (now The HDF Group, Inc.).
- Many netCDF users have made analysis, visualization, and data management software available and have made useful suggestions for enhancements to netCDF-3: www.unidata.ucar.edu/software/netcdf/credits.html

NCSA™
HDF

unidata

History of netCDF

1988 netCDF developed at Unidata

1991 netCDF 2.0 released

1996 netCDF 3.0 released

2004 netCDF 3.6.0 released

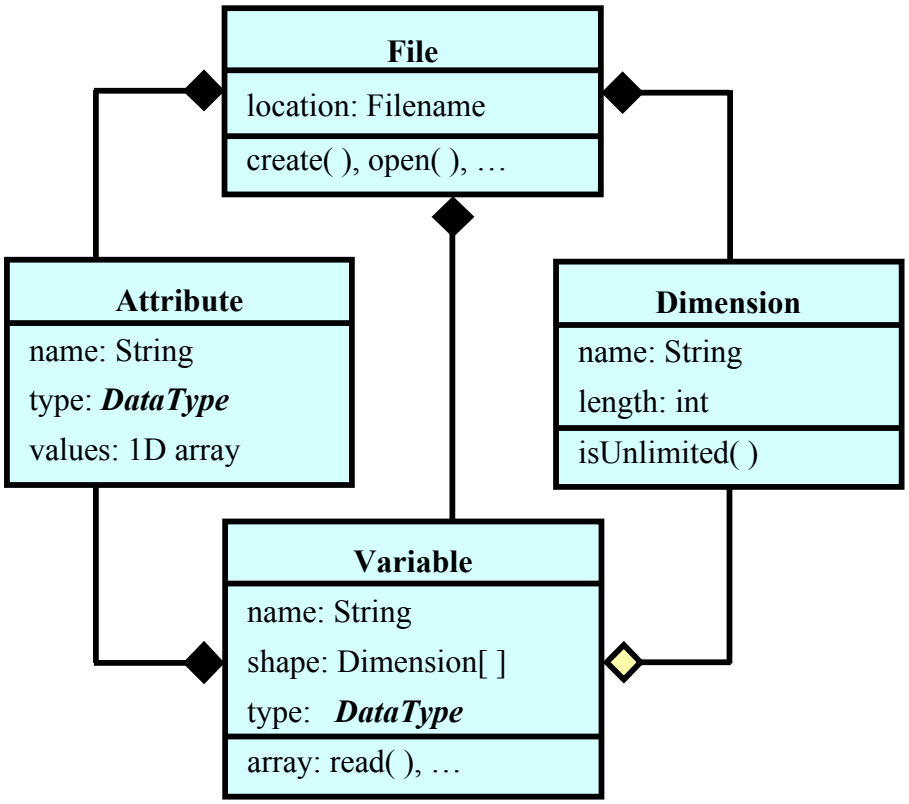
2005 netCDF 4.0 alpha released

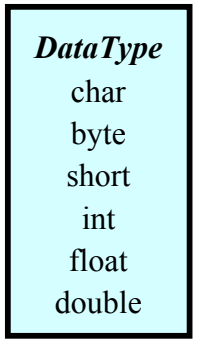
unidata

NetCDF's Niche

- Simple data model for scientific datasets
- Portable, self-describing data
- Direct access (unlike XML)
- Simple language interfaces, lots of applications:
  - C, Fortran, Java, C++, Python, Ruby, Perl
  - NCO, ncbrowse, ncview, IDV, ArcGIS, IDL, MATLAB, …
- Appendable, sharable, archivable

NetCDF-3 Data Model





Variables and attributes have one of six primitive data types.

A file has named variables, dimensions, and attributes. A variable may also have attributes. Variables may share dimensions, indicating a common grid. One dimension may be of unlimited length.

unidata

Some NetCDF-3 Limitations

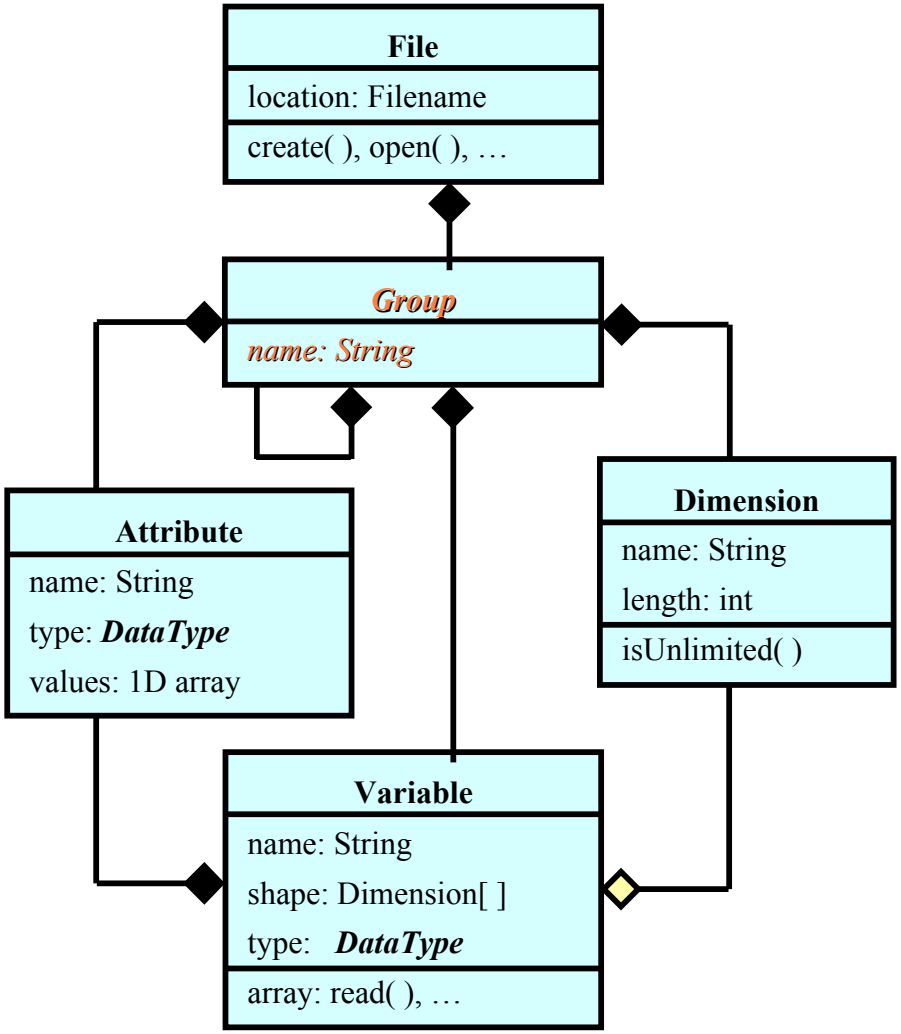
- Only one shared unlimited dimension
- No structures, just scalars and multidimensional arrays
- No strings, just arrays of characters
- Limited numeric types
- No ragged arrays or nested structures
- Only ASCII characters in names
- Changes to file schema can be expensive
- Efficient access requires reads in same order as writes
- No built-in compression
- Only serial I/O
- Flat name space limits scalability
- No querying by value or indexing for fast queries

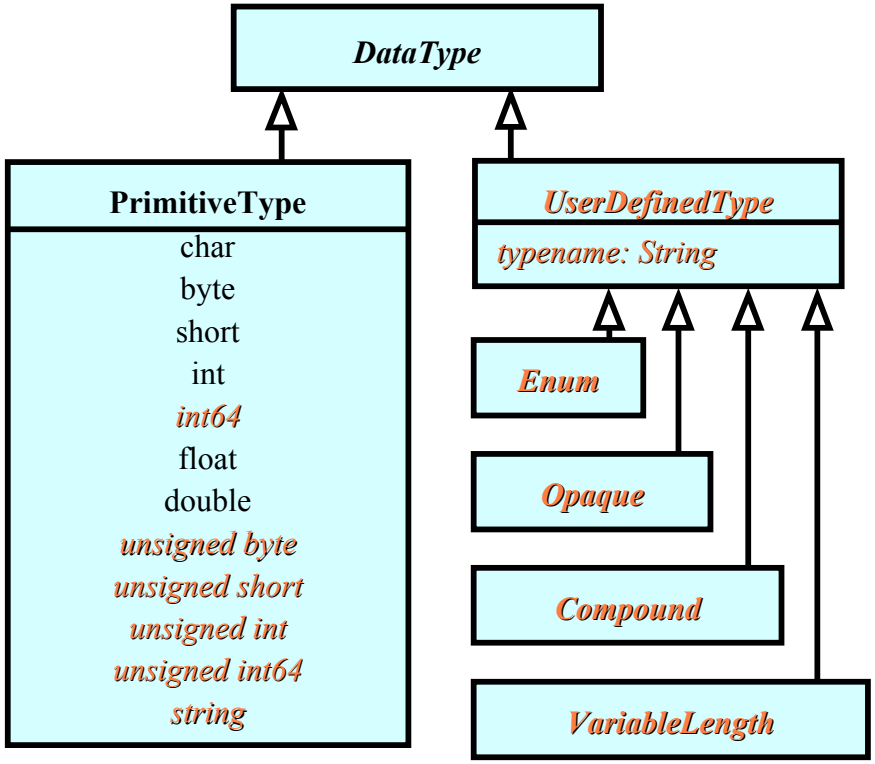
NetCDF-4 Features Address Limitations

- Multiple unlimited dimensions
- Portable structured types
- String type
- Additional numeric types
- Variable-length types for ragged arrays
- Unicode names
- Efficient dynamic schema changes
- Multidimensional tiling (chunking)
- Per variable compression
- Parallel I/O
- Nested scopes using Groups

For more details on features and their uses, see paper

NetCDF-4 Data Model



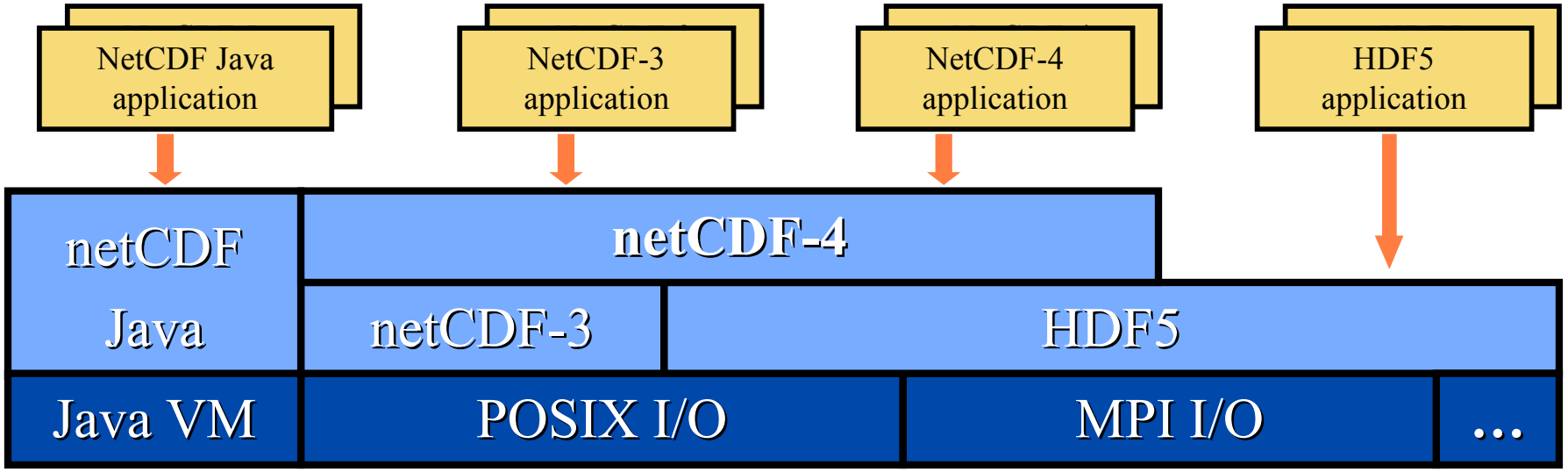


User-defined types, including compound types, may be stored with other data.

A file has a top-level unnamed group. Each group may contain one or more named subgroups, variables, dimensions, and attributes. A variable may also have attributes. Variables may share dimensions, indicating a common grid. One or more dimensions may be of unlimited length.

unidata

NetCDF-4 Architecture



- NetCDF-4 uses HDF5 for storage, high performance
  - Parallel I/O
  - Chunking for efficient access in different orders
  - Conversion using "reader makes right" approach
- Provides simple netCDF interface to subset of HDF5
- Also supports netCDF classic and 64-bit formats

unidata

Commitment to Backward Compatibility

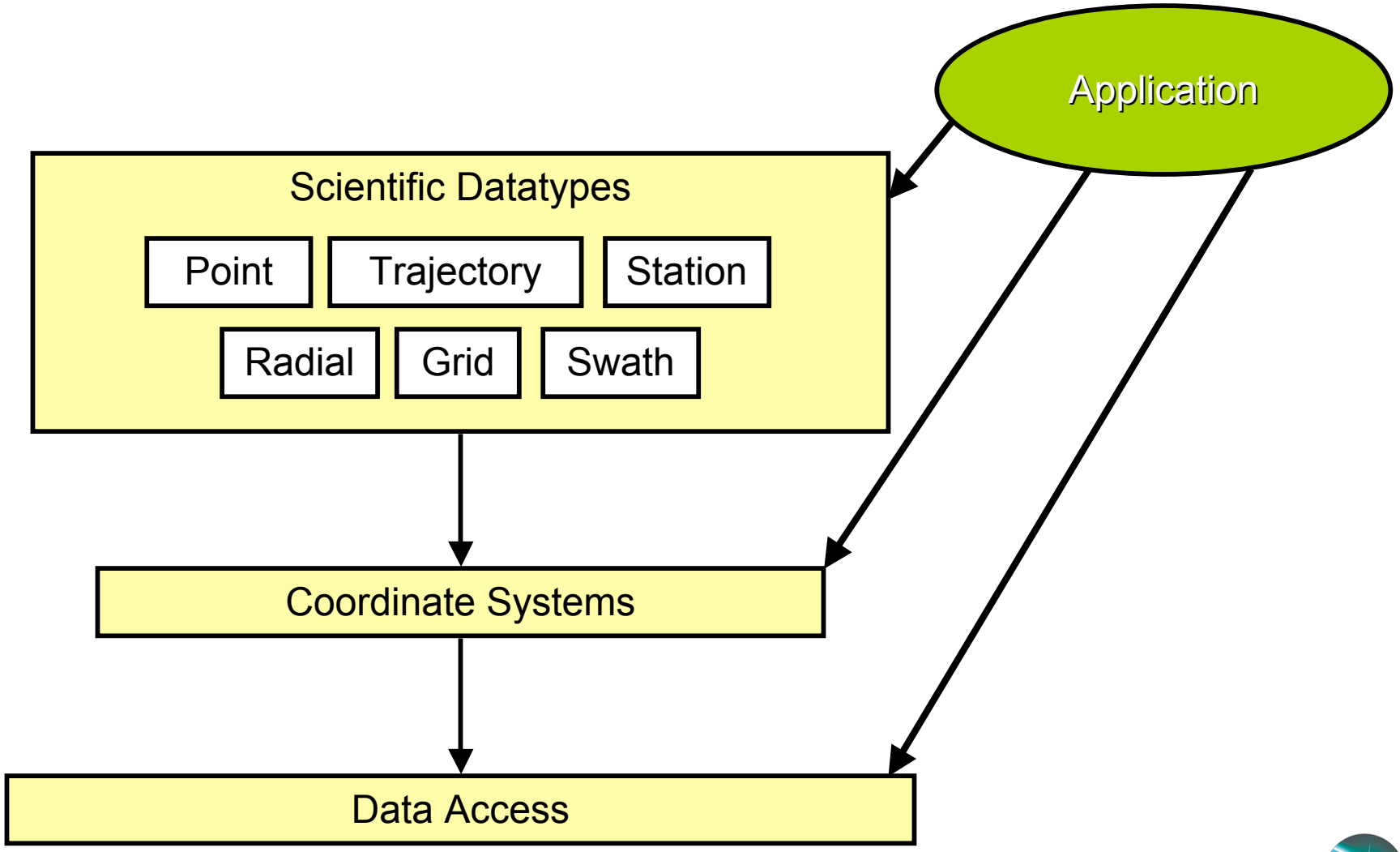
Because preserving access to archived data for future generations is sacrosanct

- NetCDF-4 provides both read and write access to **all** earlier forms of netCDF data.
- Existing C, Fortran, and Java netCDF programs will continue to work after recompiling and relinking.
- Future versions of netCDF will continue to support both data access compatibility and API compatibility.

A Common Data Access Model for Geoscience Data

- An effort to provide useful mappings among NetCDF, HDF, and OpeNDAP data abstractions
- Intended to enhance interoperability:
  - Lets scientists do science instead of data management
  - Lets data providers and application developers work more independently
  - Raises level of discourse about data objects, conventions, coordinate systems, and data management
- Demonstrated in NetCDF-Java 2.2, which can access netCDF, HDF5, OpeNDAP, GRIB1, GRIB2, NEXRAD, NIDS, DORADE, DMSP, GINI, ... data through a single interface!
- NetCDF-4.0 C interface implements data access layer

Common Data Access Model for the Geosciences



unidata

Recommendation: Adopt Cautiously

- Advanced new netCDF-4 features not yet supported by third-party programs, other language interfaces, CF conventions
- Best practices for using netCDF-4 features need to evolve
- Higher-level interfaces for coordinate systems and geoscience data objects are coming
- But … netCDF-4 writes files that are guaranteed to be readable, the netCDF classic model is easy to use, and new features may be adopted incrementally

"Every new feature is a tradeoff, between the people who could really use such a feature and the people who are just going to get overwhelmed by all the options." -- Joel Spolsky

Status and Plans

- NetCDF-4.0-alpha currently available for testing
- NetCDF-4.0
  - Awaiting HDF5 release 1.8 to finalize file format
  - Expected within a few weeks of HDF5 1.8 release
- HDF5 1.8
  - Has enhancements specifically for netCDF-4: Unicode names, dimension scales, on-the-fly numeric conversions
  - HDF5 1.8-beta expected by April 2006
- NetCDF 4.1: adds Coordinate Systems and geoscience data objects
- NetCDF 4.?: merges OPeNDAP access (pending funding)

Summary

- The current data model, APIs, and format will be supported into the indefinite future.
- The netCDF-4 release adds structs, multiple unlimited dimensions, groups, new data types, parallel I/O, and compression.
- Transition to netCDF-4's richer data model has the potential to improve interoperability and multidisciplinary use of data in the geosciences.
- For more information:
  - www.unidata.ucar.edu/software/netcdf/
  - www.unidata.ucar.edu/software/netcdf-java/
  - www.unidata.ucar.edu/staff/caron/presentations/CDM.ppt
  - support@unidata.ucar.edu

Data is Part of Our Legacy

MacKenzie Smith, Associate Director for Technology at the MIT Libraries, Project director at MIT for DSpace, a groundbreaking digital repository system

… the ephemeral nature of both data formats and storage media threatens our very ability to maintain scientific, legal, and cultural continuity, not on the scale of centuries, but considering the unrelenting pace of technological change, from one decade to the next. … And that's true not just for the obvious items like images, documents, and audio files, but also for scientific images, … and simulations. In the scientific research community, standards are emerging here and there—HDF (Hierarchical Data Format), NetCDF (network Common Data Form), FITS (Flexible Image Transport System)—but much work remains to be done to define a common cyberinfrastructure.

"Eternal Bits: How can we preserve digital files and save our collective memory?," MacKenzie Smith, *IEEE Spectrum*, July 2005

unidata