Trip Report: NetCDF-4/HDF5 Data Format Summit and NetCDF Workshop Boulder, Colorado Quincey Koziol, Larry Knox

NetCDF-4/HDF5 Data Format Summit Meeting, October 27, 2010

Ed Hartnett from Unidata proposed a meeting of netCDF and HFD5 development teams just before the annual netCDF Workshop in Boulder. Quincey and Larry were present from The HDF Group, Ed Hartnett, Russ Rew, Dennis Heimbigner, and John Caron from Unidata/UCAR. Steve Sullivan, also from Unidata came in briefly for a discussion of netCDF-4/HDF5 writing issues.

The morning was spent working through Ed's presentation of netCDF-4/HDF5 interoperability issues. This was prompted in part by the white paper on the same subject written by Elena Pourmal and Larry in September. Unfortunately Elena was unable to attend, but had also prepared a set of HDF5/netCDF-4 Interoperability Issues slides that Larry planned to present. Due to the length of the discussions of issues in Ed's presentation, Elena's slides were not presented; however, the issues raised in her presentation were well covered in Ed's.

Operability Issues and Decisions Reached

1. NetCDF-4 does not currently open HDF5 files that don't have creation order tracking enabled. This requirement may not be strictly necessary and has been relaxed for reading. It may not be a problem for writing either, and Ed will try relaxing the requirement for writing to see if it works. (However, this may expose some scalability issues in the netCDF-4 library, which Ed is aware of now)
2. Concurrent opening of HDF5 files by both libraries fails due to conflicting file close properties. The meaning of the options for these properties was discussed, and netCDF-4 will switch its default to H5F_CLOSE_WEAK, matching the HDF5 default. NetCDF-4 tests will continue to use H5F_CLOSE_SEMI to track failures to close objects.
3. NetCDF-4 does not currently open HDF5 files containing reference types. NetCDF-4 will be changed to ignore datasets with reference types. Variables (datasets) with references will not be shown by ncdump, nor presumably by API calls to get the variables in a file or group, though this was not specifically discussed. Reading attributes on such variables was not discussed either, and the presumption is that those would also be ignored.
4. NetCDF-4 crashes when reading HDF5 files with multi-dimensional arrays in attributes. NetCDF-4 allows only one-dimensional attributes. The solution proposed was to flatten multi-dimensional arrays in attributes to one dimension.
5. NetCDF-4 doesn't allow attributes for user-defined types, HDF5 does. Potentially useful (to set fill_value) but expensive for netCDF-4 to add. No decision at this time, although Ed indicated he will eventually try to accommodate the feature.

6. NetCDF-4 doesn't allow cycles in group structure, HDF5 does, therefore netCDF-4 will not open files with multiple links to an object. NetCDF-4 will open files with links, will keep the first link to an object that it finds and discard any subsequently discovered links, forcing a tree-like structure and making each object visible in only one location.

Other Suggestions and Questions

1. NASA wants a document to describe interoperability (for data providers). Perhaps Ed and Elena can produce it cooperatively.
2. GIP would like a similar guide for users – probably less detailed.
3. There's a strong desire to cooperate on HDF4 interoperability. We may want to retire the netCDF interface in the HDF4 library, but then the netCDF library would have to add the ability to write to HDF4 files (currently it is only reading them).
4. OpenDAP collaboration. Should we test that netCDF-Java and an HDF5 OpenDAP server return the same answers about an HDF5 file? (Probably they are different)
5. HDF-EOS. NetCDF-4 needs an augmentation tool for HDFEOS5. Ed would like to work with Kent for a day, thinking this should be enough time to be helpful.
6. The netCDF team would like to leverage our perf. regression tests, particularly on HPC platforms, and would like to collaborate with us on this.
7. Serious dimension scale issue – NetCDF-4 allows 2-dimensional scales. How hard would it be for HDF5 to allow these? Solution needed from HDF5.
8. The lack of shared dimensions in HDF5 is a real problem for netCDF-4. We really need to tune up the HDF5 dimension scales model.
9. Will the C netCDF-4 library add support for NcML? When? The idea was mentioned at the workshop as being good and useful, but probably not within the next year unless user interest (or funding?) raises the priority.
10. How can one generate an NcML file? Is it used widely or is it new to most of the scientists? NcML was described as a dialect of XML for representing netCDF data, for virtual netCDF files, and for virtual aggregations. It can be generated from classic netCDF data by ncdump (with –x option). The netCDF-Java Tools UI utility can also convert netCDF and CDL data to NcML . Since netCDF-Java can read HDF5 files, maybe it can also convert HDF5 files, without the references and links that it doesn't support? My impression is that NcML is essentially CDL with xml tags, so even if not widely used yet there should be netCDF users for whom it will not be totally unfamiliar.
11. What are Ed's plans for allowing file ids along with the "long" variable names? This didn't come up in the Tuesday meeting and Ed was somewhat evasive when asked about this one. In the past he has sounded like it wouldn't be difficult to do. I think it's a lower priority for him, but given some assistance, particularly in testing it seemed to be a possibility.

In the afternoon Cecilia DeLuca gave a talk about NOAA's Global Interoperability Program. The scale of this program is much broader and less focused than NetCDF-4/HDF5 interoperability and probably doesn't directly affect HDF5.

More time was spent discussing interoperability issues, mostly from the netCDF-4 perspective. At Ed's invitation, Larry gave a short presentation of HDF5 daily testing.

The last talk of the day was delivered by Carlos Maltzahn who was accompanied by three of his graduate students from the University of California at Santa Cruz. They are working on a data storage project to be built on CEPH, a file system that is now part of the Linux kernel. The concept of the project is to move database capabilities and query optimization into the file system.

NetCDF-4 Annual Workshop, October 28-29, 2010

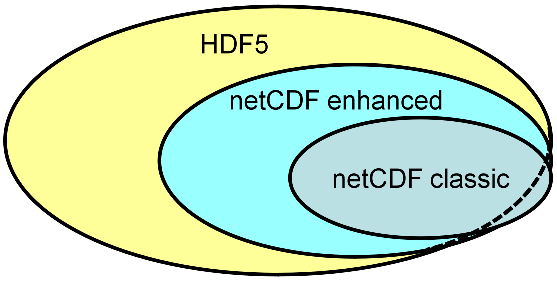
Quincey stayed to deliver Elena's introductory and advanced talks at the workshop on Thursday and Friday. The program was a fairly comprehensive overview of NetCDF and some of the applications that use it. The schedule bogged down on Thursday, so Quincey combined the two talks on Friday morning. With all the talk of interoperability it may be worthwhile to investigate whether there are highly useful HDF5 calls that could be added to a netCDF-4 application, and if so consider creating some examples as to how to do that.

The talks also included netCDF-4/HDF5 interoperability and near-term and long-term netCDF plans. The intentions to ignore references and extra links to groups, which would be consistent with netCDF-Java were described in future plans. These changes will not be in the imminent (hopefully by the end of the year) release of netCDF 4.1.2, but in netCDF 4.2, planned for early 2011.

The slide below caught my attention. The ideal is that the three circles should be concentric, but the reality is that there is a small but essential part of netCDF that is outside HDF5, creating significant obstacles to interoperability. We should take a look at whether we might be able to expand HDF5 to include the bulge. (The main problem is the lack of shared dataspaces in HDF5)

The Enhanced NetCDF Data Model

- Additions to classic netCDF data model
- Still a subset of HDF5 data model
- Made possible by adding a few things to HDF5 so netCDF classic data model could fit within it
- Criteria for additions: handling identified classic limitations, simplicity
- Issue: is enhanced netCDF data model the right balance of simplicity and power?


HDF5
netCDF enhanced
netCDF classic