Status of LOFAR Data in HDF5 Format

Anastasia Alexov[1,2], Pim Schellart[3], Sander ter Veen[3] and the LOFAR ICD Group

1: University of Amsterdam (UvA) [a.alexov@uva.nl], 2: Netherlands Institute for Radio Astronomy (ASTRON); 3: Radboud University Nijmegen

WWW.LOFAR.ORG

The Low Frequency Array (LOFAR) project is solving the challenge of data size and complexity using the Hierarchical Data Format, version 5 (HDF5). Most of LOFAR's standard data products will be stored using the HDF5 format; the Beam-Formed (aka Pulsar) and Transient Buffer Board (TBB) Time Series Data have transitioned from project-specific binary to HDF5 format. It supports an unlimited variety of datatypes, and is designed for flexible and efficient I/O and for high volume and complex data. HDF5 is portable, extensible, and parallelizable, allowing applications to evolve in their use of HDF5. We report on our effort to pave the way towards new astronomical data encapsulation using HDF5, which can be used by future ground and space projects. The LOFAR project has formed a collaboration with NRAO, the VOA and the HDF Group to obtain funding for a full-time staff member to work on documenting and developing standards for astronomical data written in HDF5. We hope our effort will enhance HDF5 visibility and usage within the community, specifically for SKA Pathfinders and other major new radio telescopes such as LOFAR, EVLA, ALMA, ASKAP, MeerKAT, MWA, LWA, and eMERLIN.

The LOFAR Radio Telescope

- The "LOw Frequency ARray" (LOFAR)
- Currently, the largest radio telescope in the world
- 48 stations (40 NL, 8 International); complete by end of 2011 *(Figure 1 on right shows the center stations in NL)*
- Baselines from 1 - 1500 km; ultimately achieve sub-arcsecond resolution over much of the band
- Low Band Antenna bandpass: 30-80 MHz
- High Band Antenna bandpass: 120-240 MHz
- Total Recordable Bandwidth: 48MHz; 100MHz TBBs
- Data Correlation: IBM Blue Gene/P supercomputer, Groningen, NL *(see photo in Figure 2 on right)*
- Offline processing cluster has 100 nodes, each with: 24 cores, 64GB RAM, 21TB storage
- Long Term Archive (LTA) has: 2.2PB disk, 5PB tape
- Access to 22,600 cores via BigGrid and JUROPA

The LOFAR "Superterp", Exloo, NL (Fig. 1)

The Blue Gene/P supercomputer at Groningen (Fig. 2)

LOFAR Data in HDF5: Solves Variety, Complexity & Size Issues

Time Series data share similar issues to all datasets produced by LOFAR observations -- they vary tremendously in type, size and complexity. Instead of using many different file formats to encapsulate LOFAR data, the Hierarchical Data Format, version 5 (HDF5) was chosen as a viable solution for potentially massive LOFAR data products.

LOFAR data statistics:

- Data rates up to 8GB/sec
- File sizes, 10s to 100s TB
- Datasets ranging from 1 up to 6 dimensions
- Single dataset can be spread over multiple disks
- Data writing must be multi-threaded and parallelizable
- Most astronomical data containers (CASA MS or FITS, or ~30 pulsar data formats) could not meet all our needs

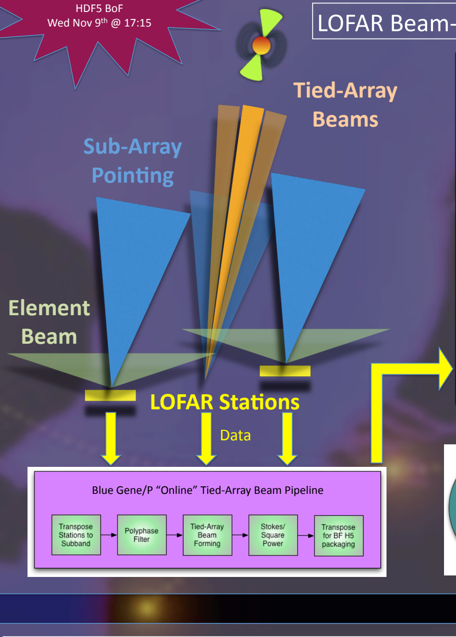
Why use HDF5?

- Format for managing and storing large and complex scientific data
- Unlimited variety of datatypes
- No inherent file size limitations
- Parallel reading and writing (MPI)
- Runs on massively parallel/ distributed systems
- Built-in compression
- Library API in C/C++/Java/Fortran
- Self-describing and portable
- Free & has 20+ year history @NASA

Hear more at: HDF5 BoF Wed Nov 9th @ 17:15

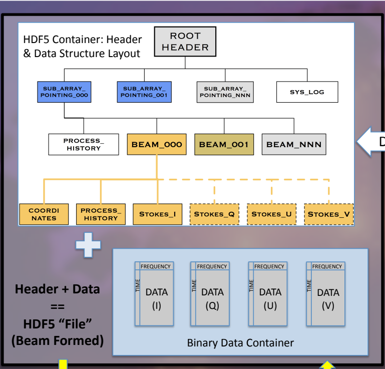
LOFAR Beam-Formed Time Series Data Format Specifications in HDF5

www.hdfgroup.org



Blue Gene/P "Online" Tied-Array Beam Pipeline

Transpose Stations to Subband → Polyphase Filter → Tied-Array Beam Forming → Stokes/ Square Power → Transpose for BF H5 packaging

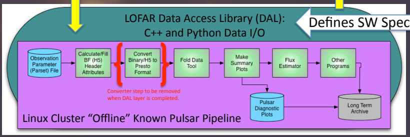


HDF5 Container: Header & Data Structure Layout

Binary Data Container

Header + Data == HDF5 "File" (Beam Formed)

Describes Layout



LOFAR Data Access Library (DAL): C++ and Python Data I/O

Defines SW Specification

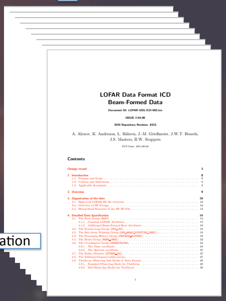
Observation Parameter (Parset) File → Calculate/Fill BF (H5) Header Attributes → Convert Binary/H5 to Presto Format → Fold Data Tool → Make Summary Plots → Flux Estimator → Other Programs → Long Term Archive

Make Summary Plots → Pulsar Diagnostic Plots → Long Term Archive

Converter step to be removed when DAL layer is completed.

Linux Cluster "Offline" Known Pulsar Pipeline

For the LOFAR project, a data product is defined within the context of HDF5. HDF5 allows for storage, not only of the data, but for the associated and related meta-data describing the data's contents, conditions of observations, logs, etc.. As an "all-in-one" wrapper, the HDF5 format simplifies the management of complex datasets.



LOFAR Data Format ICD
Beam-Formed Data

Document ID: LOFAR-USG-ICD-003.tex

ISSUE 2.04.08

SVN Repository Revision: 8333

A. Alexov, K. Anderson, L. Bähren, J.-M. Grießmeier, J.W.T. Hessels, J.S. Masters, B.W. Stappers

SVN Date: 2011-09-02

Contents

1

LOFAR Interface Control Documents[1] (ICDs) provide detailed descriptions of all expected LOFAR Data Products in **HDF5** format:

- **Beam-Formed (BF) Data** ***[shown on the left]***
- **Transient Buffer Board (TBB) Time Series** ***[bottom left of poster]***
- **Radio Sky Image Cubes**
- **Dynamic Spectrum Data**
- **Rotation Measure (RM) Synthesis Cubes**
- **UV Visibility Data[2]**
- **Near-field Images[2]**

LOFAR project is developing the C++ **Data Access Library (DAL)**, which provides full scope constructors for creating and accessing LOFAR data products, using the ICD specifications. TBB and Beam-Formed have been implemented.

[1] *Documents available from the **LOFAR Wiki ("Data Products" link), http://lus.lofar.org/wiki***

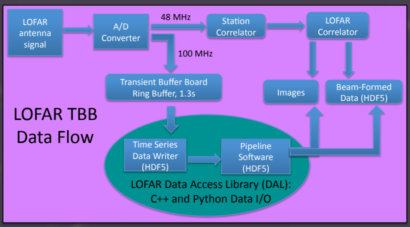
[2] *UV Visibility and Near Field Imaging are under development*

Transient Buffer Board (TBB) Data: look-back into the "past"

Each LOFAR dipole antenna has a ring buffer, transient buffer board, that can store up to 1.3 seconds of raw time series data sampled every 5ns (LOFAR's highest time resolution). LOFAR acts as a "time machine" – a transient event triggers the buffered data to be written to HDF5, providing up to 1.3 seconds of historical information at full temporal and spatial resolution.

TBB HDF5 Data Stats:

- 100 MHz of bandwidth
- 1.3 sec of look-back time
- 40MB - 48GB dataset/station
- 2GB - 2.25TB dataset sizes
- Data maps easily to HDF5
- 2012 Upgrade: 5 sec look-back
- 2012 Upgrade: 8.6 TB dataset



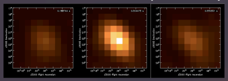
LOFAR TBB Data Flow

LOFAR antenna signal → A/D Converter → (48 MHz) Station Correlator → LOFAR Correlator → Images; Beam-Formed Data (HDF5)

A/D Converter → (100 MHz) Transient Buffer Board Ring Buffer, 1.3s → Time Series Data Writer (HDF5) → Pipeline Software (HDF5) → Images; Beam-Formed Data (HDF5)

LOFAR Data Access Library (DAL): C++ and Python Data I/O

TBB's are used to study: air showers produced by cosmic rays, lightning on Earth and other planetary bodies within our own solar system, and unknown sub-second timescale astronomical transients.



Crab Giant Pulse (@center), as seen by the TBB boards at 0.015 sec sampling time

Summary, Collaborations And Future Considerations

HDF5-friendly Tools: **IDL, VisIt, HDFView, h5py, PyTables, DAL** (LOFAR), **PyDAL** (LOFAR)

LOFAR has started work on the next generation of its C++ **Data Access Library (DAL + PyDAL)**, based on current time series data use cases and use patterns, with the goal of speed-up, simplification of use and better python bindings. We also intend to expand the DAL for use with additional LOFAR data formats and to be able write converters to/from FITS, the Casa Tables Measurement Sets and HDF5. Improvements coming in 2012!

LOFAR has teamed up with NRAO, the IVOA and The HDF Group in requesting funding for defining and developing astronomical data standards in HDF5, based on LOFARs ICD and DAL ground work. This effort is nicknamed "**AstroHDF**". Future work also involves developing converts to/from HDF5/FITS and collaborations on interfaces in **DS9** and **VisIt** for HDF5 (WCS-compliant) astronomical data.

We believe that the HDF5 format is well-suited for complex and large astronomical data. For some of LOFAR's data format challenges, HDF5 is the only viable solution. In HDF5, LOFAR can easily map 10s of TB of complex, 6-dimensional data structures with intricate header and logging information throughout the data-files. We feel that HDF5 can meet the data needs and demands of future projects such as LSST and the SKA, where size and complexity will be even greater of an issue than faced by LOFAR.

Collaborative mailing list info: nextgen-astrodata@astron.nl
Sign up: majordomo@astron.nl w/message body "subscribed nextgen-astrodata"

Hear more at: HDF5 BoF Wed Nov 9th @ 17:15

LOFAR

UNIVERSITY OF AMSTERDAM

NWO

ASTRON