LOFAR and HDF5: Towards a new Radio Standard

(see poster)

Anastasia Alexov

On behalf of the LOFAR Data Formats Group

IDIA 2011 (Lightning Talk)

LOFAR and HDF5: Towards the Next Generation Astronomical Data Standard

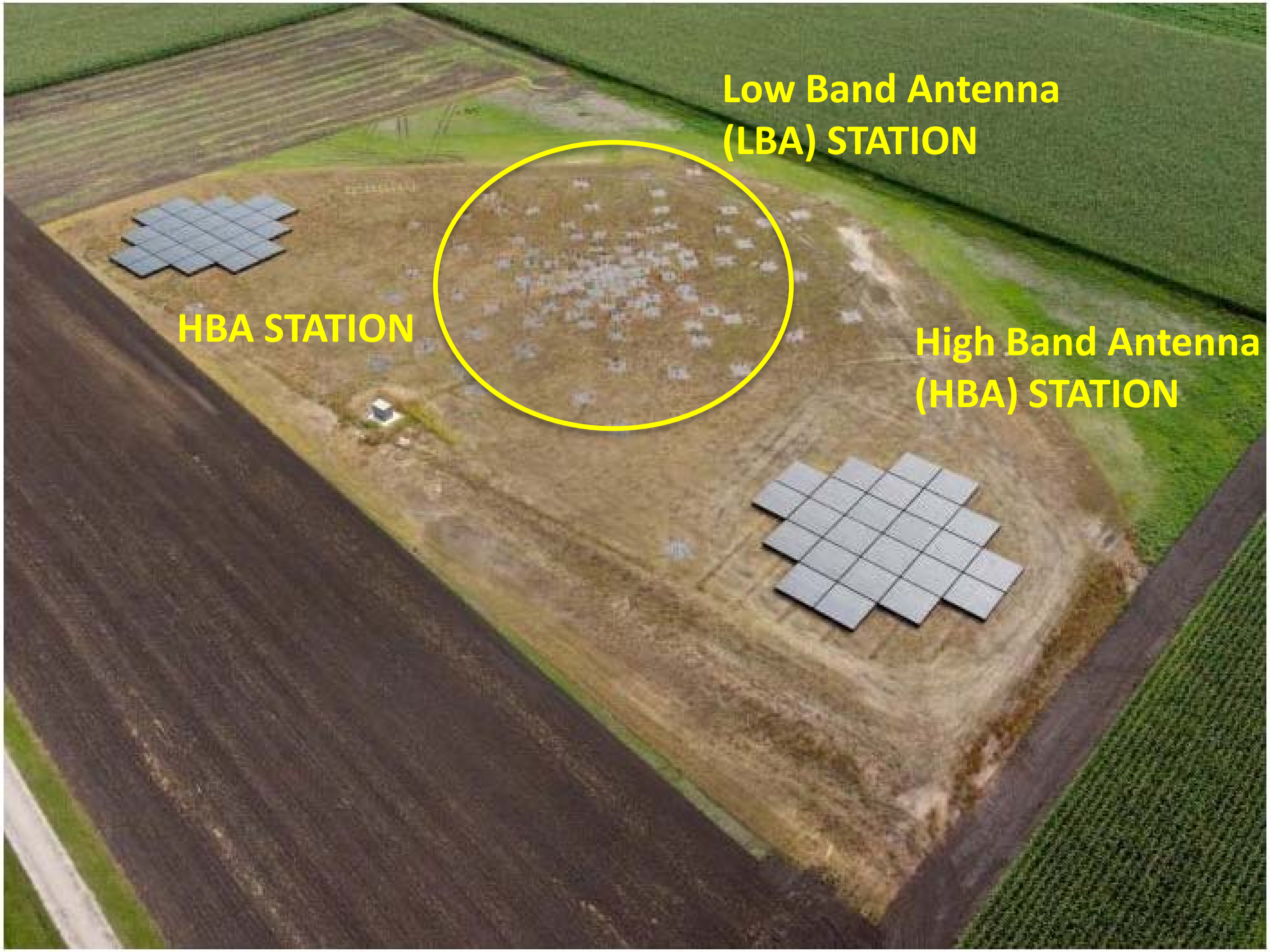
(see poster)

Anastasia Alexov

On behalf of the LOFAR Data Formats Group

IDIA 2011 (Lightning Talk)

The "LOw Frequency Array"
LOFAR



LOFAR

Currently Operational:
- 36 stations (NL)
- 5 stations (EUR)
- 48 MHz bandwidth
- Frequency Range
  - 30-80 MHz (Low Band Antenna)
  - 120-240 MHz (High Band Antenna)
- 8+ simultaneous beams
- Baselines from 1 - 1500 km
- Data Correlation: IBM Blue Gene/P supercomputer, Groningen, NL
- Offline processing cluster has 100 nodes, each with: 24 cores, 64GB RAM, 21TB
- Long Term Archive (LTA) has: 2.2PB disk, 5PB tape
- Access to 22,600 cores via BigGrid and JUROPA
- 0.76 kHz (1 sec) spectral resolution
- 5.1 nano-second time resolution

Core Stations

Completed 2011:
40 stations (NL)
8+ stations (EUR)

Data Variety, Complexity and Size!

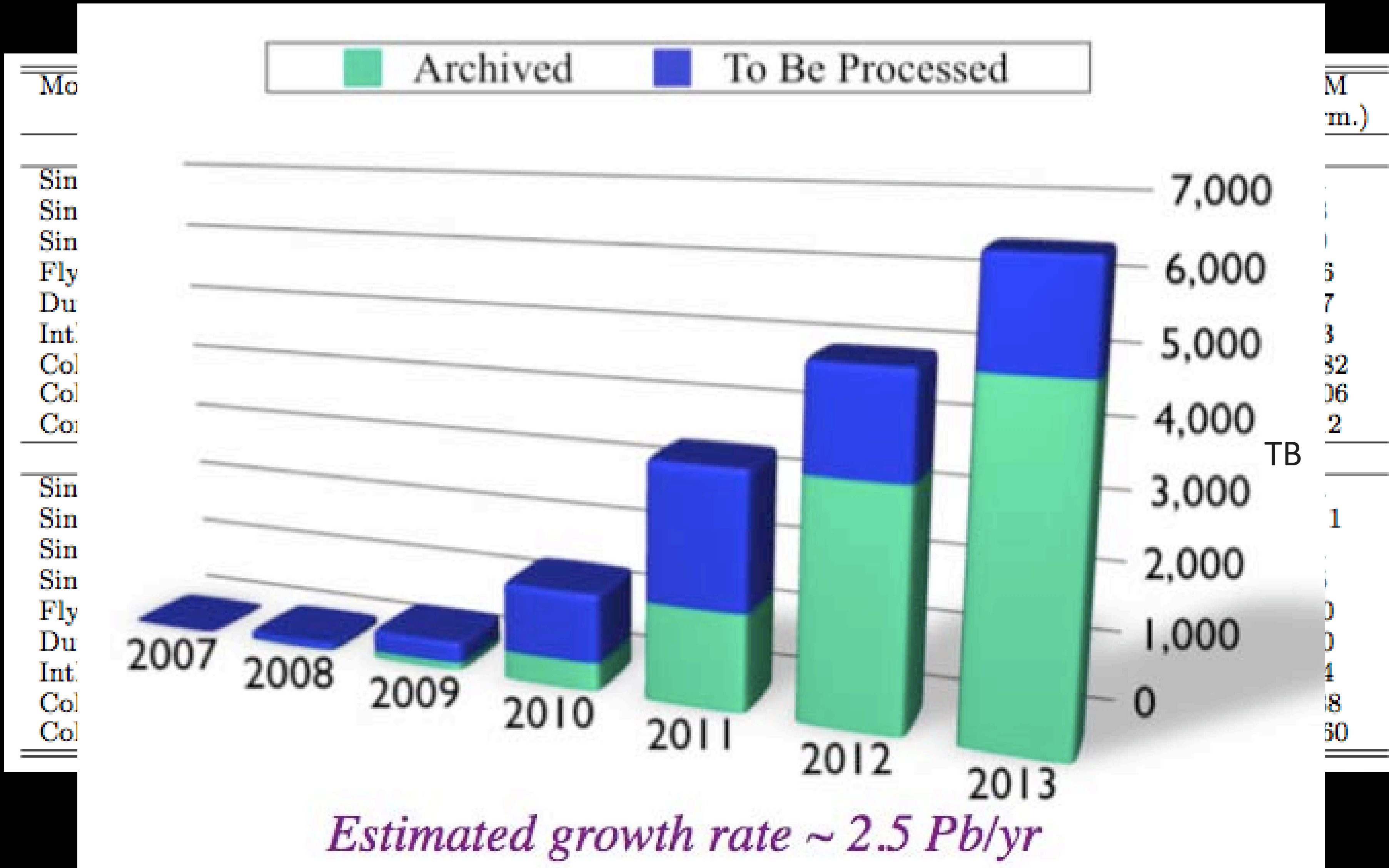
- LOFAR has many observing modes [Imaging/Visibility Data, Beam-Forming (BF)/Time-Series, Transient Buffer Board (TBB) dumps, Rotation Measure (RM) Synthesis, Dynamic Spectra, etc]
- Many different observing modes create data diversity/variety [6 basic LOFAR data types]
- Using 30+ LOFAR stations creates enormous data rates and sizes [max ~31TB/hour]

LOFAR Data Dimensionality

Data product	Quantity	Array shape
TBB time-series	$s_j(t)$	1-dim
TBB spectral-data	$\widetilde{s}_j(\nu)$	1-dim
Beam-formed data	$\widetilde{S}(p, \nu, \mathrm{Dec}, \mathrm{RA})$	3-dim
All-sky dynamic spectrum	$I(p, \nu, t)$	3-dim
Visibility data	$V(p, \nu, t, B_{12})$	4-dim
Image hypercube	$I(p, t, \nu, \vec{\rho})$	6-dim
Radio sky image	$I(p, \nu, Dec, RA)$	4-dim
CR image cube	$I(p, t, \nu, r, \mathrm{El}, \mathrm{Az})$	6-dim
CR image cube	$I(p, t, \nu, \xi_3, \xi_2, \xi_1)$	6-dim
RM Synthesis cube	$\mathrm{DF}(p, \mathrm{Dec}, \mathrm{RA}, \phi)$	4-dim
RM Synthesis map	$\mathrm{RM}(\mathrm{Dec}, \mathrm{RA})$	2-dim

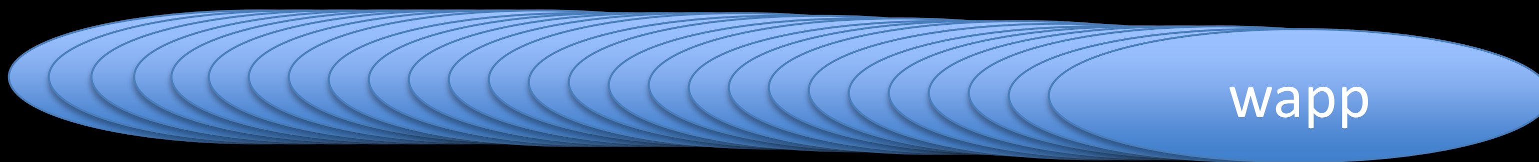
Polarization,
Time,
Frequency,
Position 1,
Position 2,
Position 3

LOFAR Data Sizes



Astronomical Data "Containers"
(non-comprehensive list)

- Binary (especially for Time-Series data for each telescope/instrument):



- FITS (all wavelengths)
- CASA [casacore] (Radio)
- MBFITS (Multi-Beam FITS for Radio)
- And many OTHERS (usually for ONE type of data)

…As a result:

- Software data I/O becomes a mess!
- Tools have to be adapted per data container

Why LOFAR chose (yet) another data format: Hierarchical Data Format, version 5 (HDF5)

Question: Can only ONE of the astronomical formats (like ~~FITS~~ or ~~CASA~~) do ALL these things?

- HDF5 is a data model, library, and file format for storing and managing large and complex scientific data (images, N-D arrays, tables, metadata).
- It supports an unlimited variety of datatypes, and is designed for flexible and efficient I/O and for high volume and complex data.
- Self-describing and portable to a diversity of computational environments
- No inherent size limitations
- C, C++, Java, Fortran 90 interfaces
- Can be run on single node or massively parallel/distributed systems
- Built-in compression (GNU zlib, but can be replaced with others)
- Parallel reading and writing (via MPI-I/O)
- Partial I/O: “Chunked” (tiled) data for faster access
- Free and in use for 20+ years by NASA and other projects
- Inspection and visualization tools exist (HDFView + command line tools, VisIt + pluggin, PyTables, h5py, MATLAB, IDL)

LOFAR Data Interface Control Documents (ICDs)

LOFAR Data Format ICD
Rotation Measure Synthesis Cubes

LOFAR Data Format ICD
Visibility Data

LOFAR Data Format ICD
Dynamic Spectrum Data

LOFAR Data Format ICD
TBB Time-Series Data

LOFAR Data Format ICD
Beam-Formed Data

LOFAR Data Format ICD
Radio Sky Image Cubes

LOFAR Data Format ICD
Representations of World Coordinates

Document ID: LOFAR-USG-ICD-002

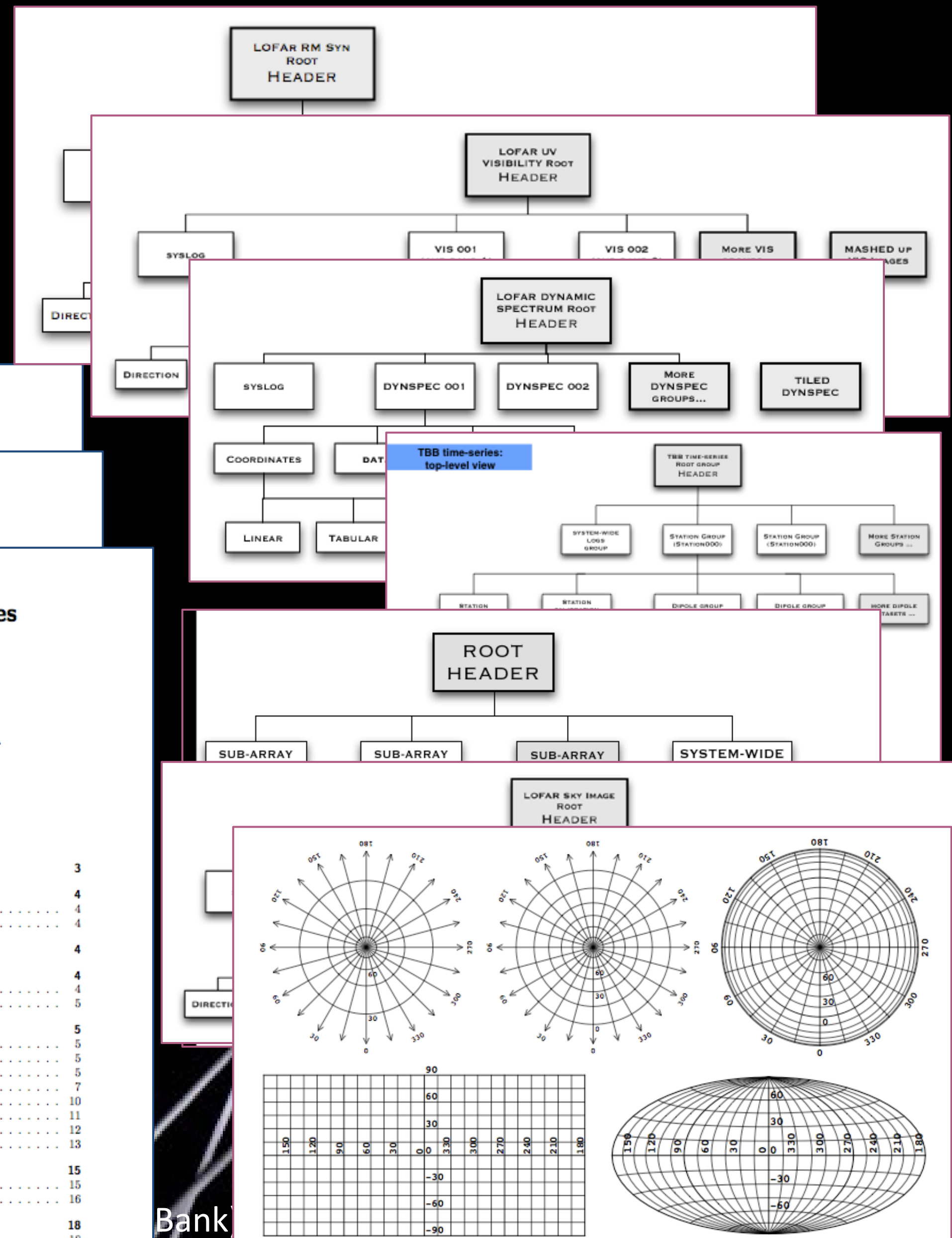
Version 2.00.00

SVN Repository Revision: 6137

L. Bähren, A. Alexov, K. Anderson, J.-M. Grießmeier

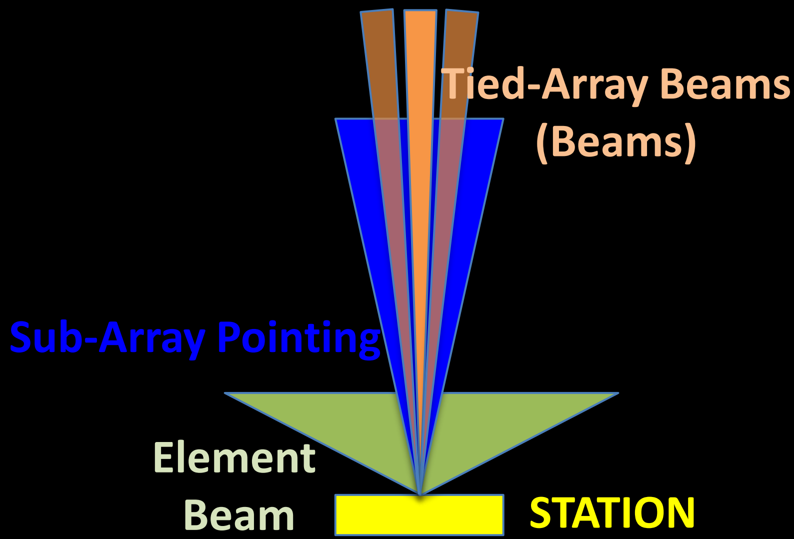
SVN Date: 2010-10-27

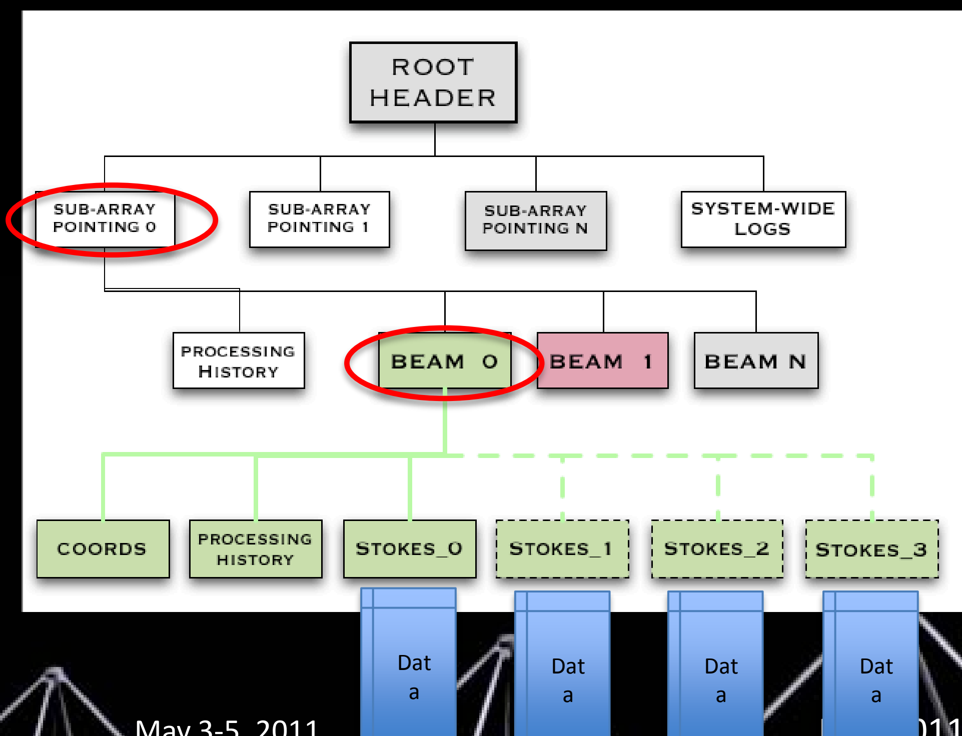
Contents



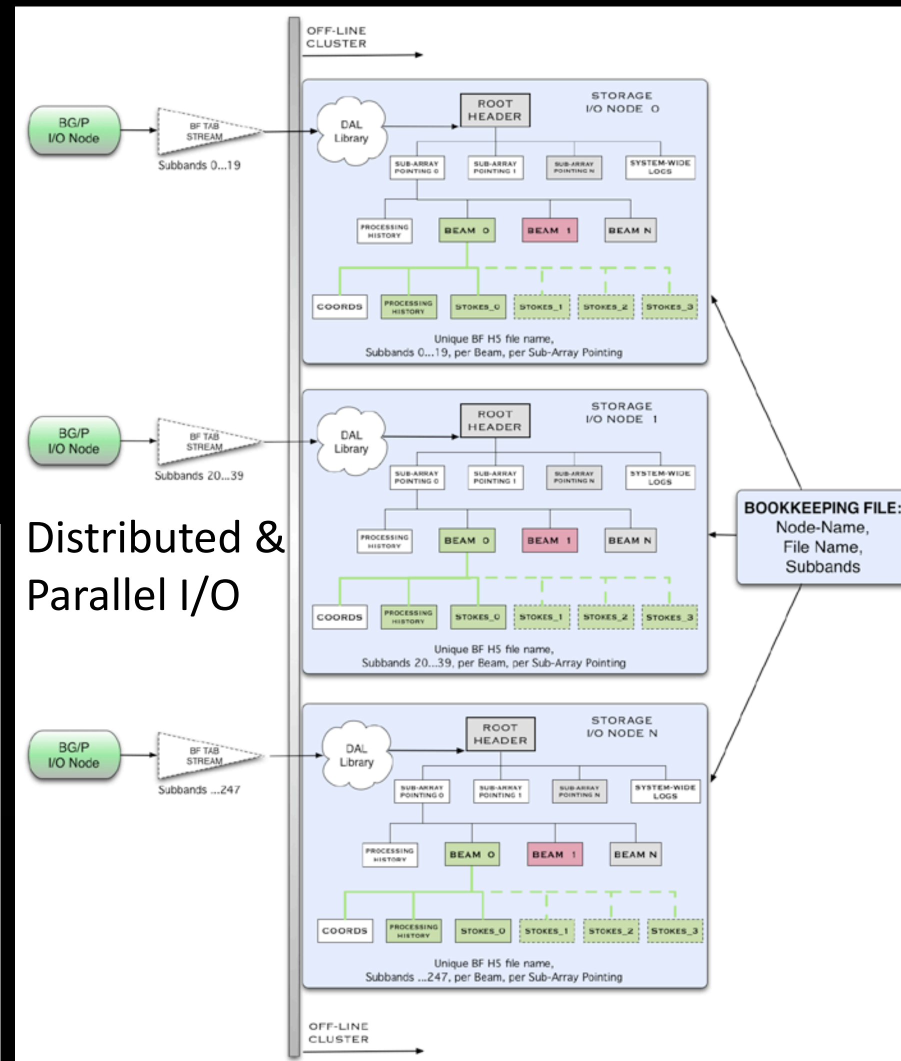
Bank)

Beam-Formed Data Format in HDF5:





Distributed &
Parallel I/O



LOFAR Data Access Software

- LOFAR User Software (LUS) available in SVN repository (cmake build):
  - LOFAR tools, pipelines, etc. (C, C++, Python, etc)

- C++ Data Access Layer (DAL) Library (intermediate layer on top of HDF5)
- DAL Python wrapper (PyDAL)

  https://github.com/nextgen-astrodata/DAL
- C++ Classes are based on LOFAR data format ICDs [Beam Formed, Sky Image, Dynamic Spectra, Transient Buffer Board]

- Work in progress:
  - HDF5 Data I/O benchmarking
  - Choosing optimum HDF5 data containers [dim, cache, chunk] (adjust ICDs as needed)
  - LOFAR HDF5 Data writers
  - Plan on visualization tool: pluggin for VisIt
  - Plan on H5 Sky Cube -> FITS converter for DS9

Datasets of the Future… (in HDF5)

- Future telescopes have similar challenges:
  - Radio: EVLA, ALMA, ASKAP, MeerKAT, MWA, LWA, eMERLIN and SKA!
  - Non-Radio: Pan-Starrs, LSST, TMT, GMT, ELT
- MeerKat project is writing HDF5 using python (benchmarking PyTables vs h5py); evaluating LOFAR ICDs
- Simulation community uses HDF5 (GADGET, ENZO, FLASH); HDF5-iRODS Grid project
- Collaborations needed to expand HDF5-usage and tool-set in astronomy; discuss on moderated mailing list: nextgen-astrodata@astron.nl

  Email to: majordomo@astron.nl

  Text in message body: subscribe nextgen-astrodata

- Don't be fooled into thinking binary is the only solution – issues with long-term maintenance and lack of astronomy tool-sets
- Time is ripe to solve this issue across wavelengths and projects; HDF5 is mature and used extensively in science
- This is NOT just a "Radio-problem", it's an astronomical problem!