Next Generation Data Management: From LOFAR to the SKA

Innovations in Data-Intensive Astronomy

May 5, 2011

Michael Wise
ASTRON / LOFAR / UvA

Next Generation Data Management: From LOFAR to the SKA

Overview of LOFAR

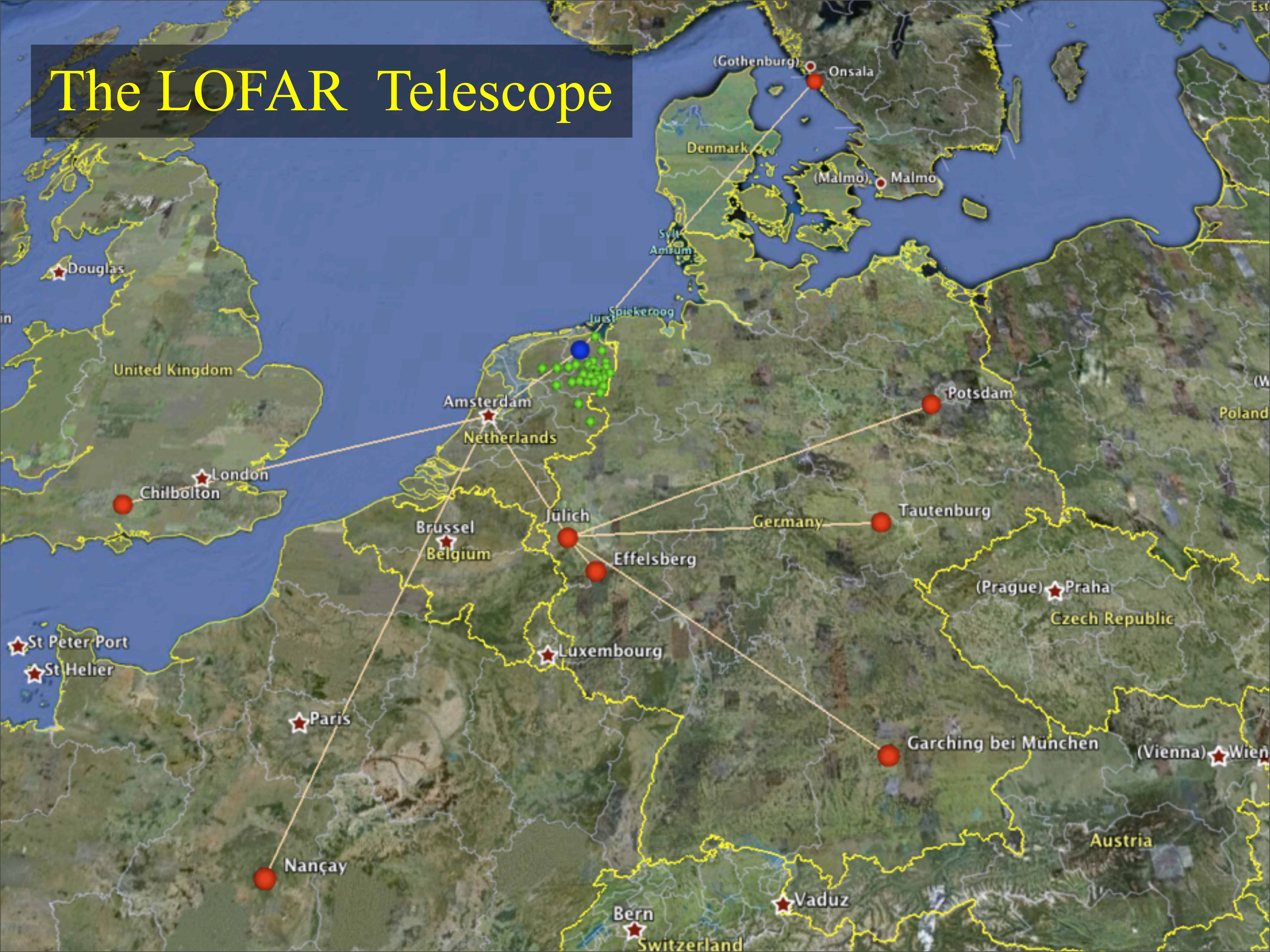
Science drivers and pipelines

Data flow and bottlenecks

Data products and access

Modern archives

The LOFAR Telescope

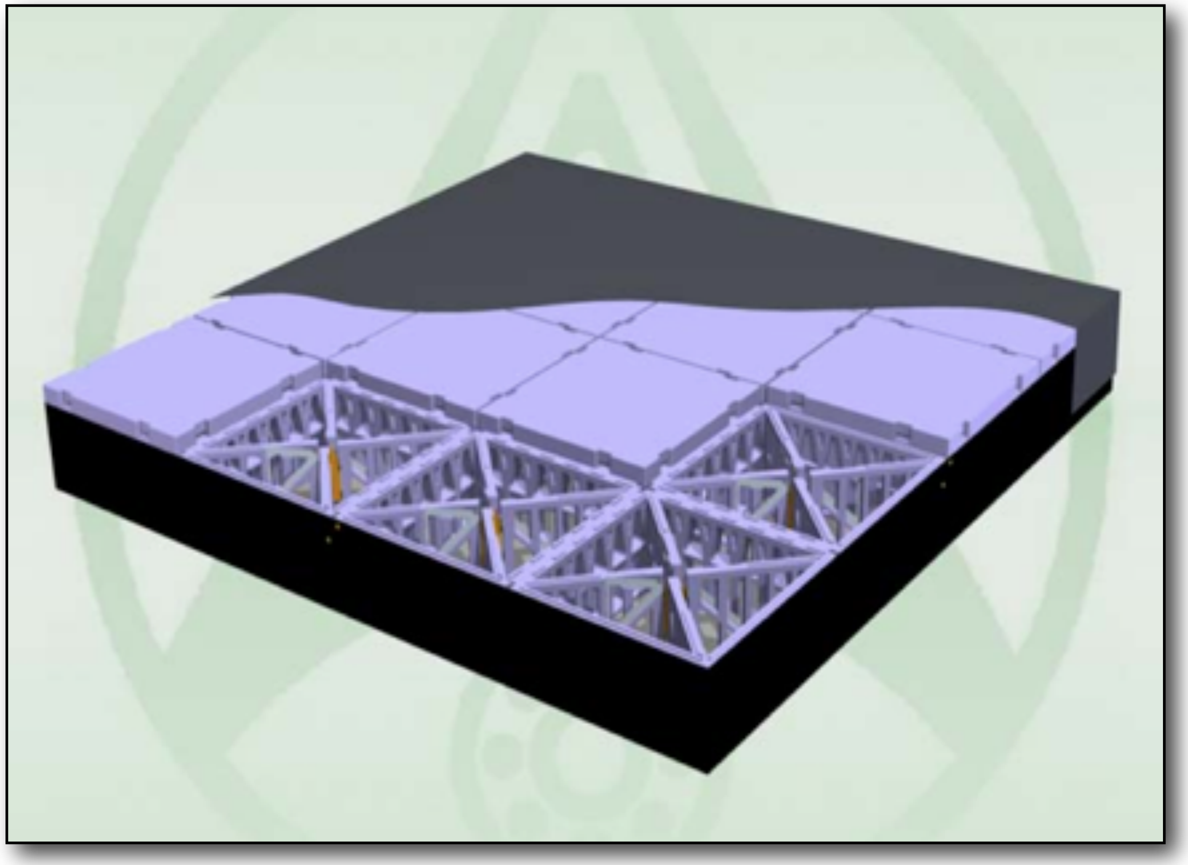



LOFAR Superterp
June 2010

LOFAR Antennas

High band tiles:120 – 240 MHz
96 tiles/station, 4x4 antennas/tile

Low band antenna: 30 – 80 MHz
48/96 antennas per station

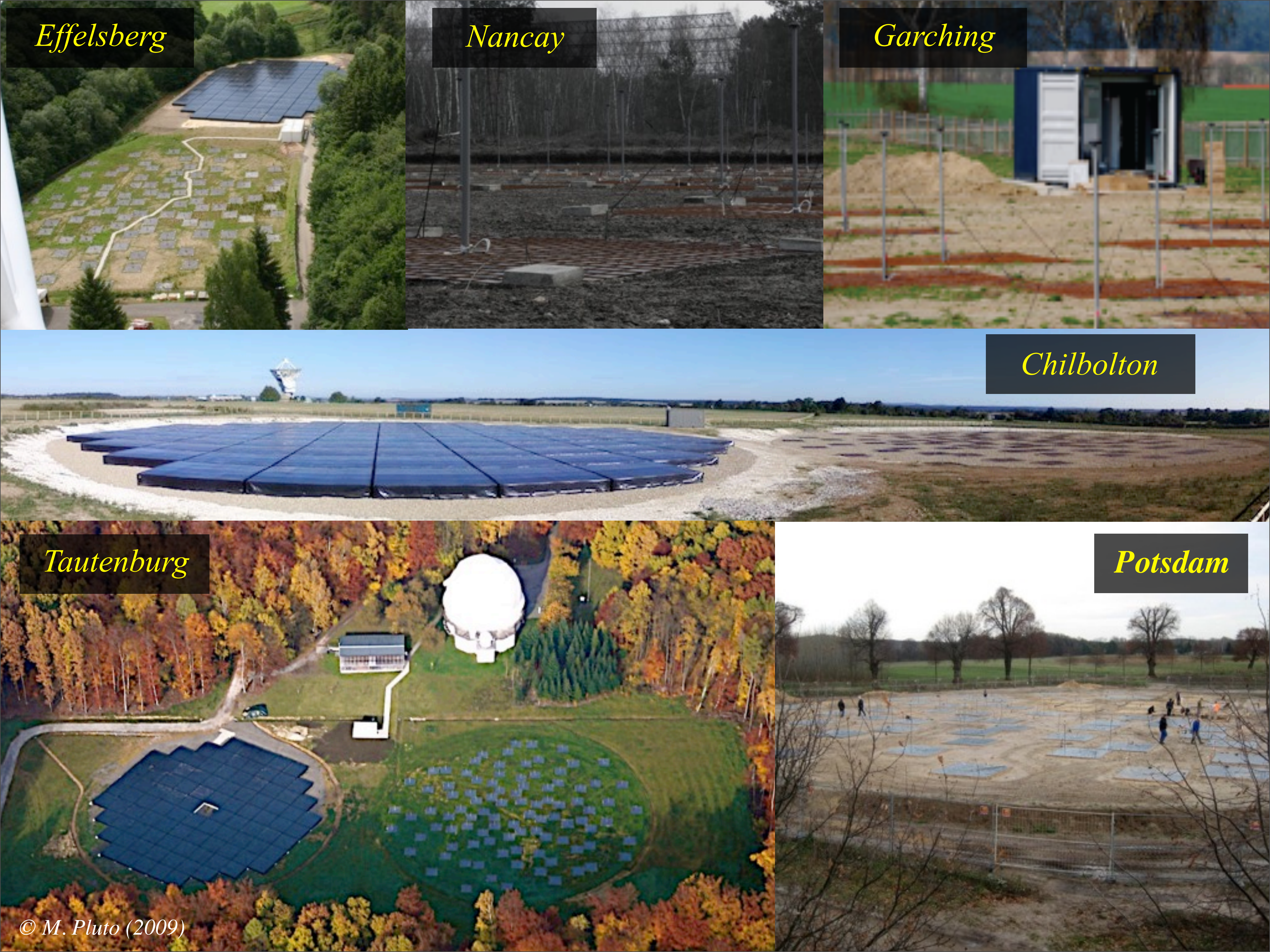


- 40 NL + 8 EU stations of dipoles
- Replace big dishes by many cheap dipoles
- No moving parts: electronic beam steering
- Flexible digital beam forming

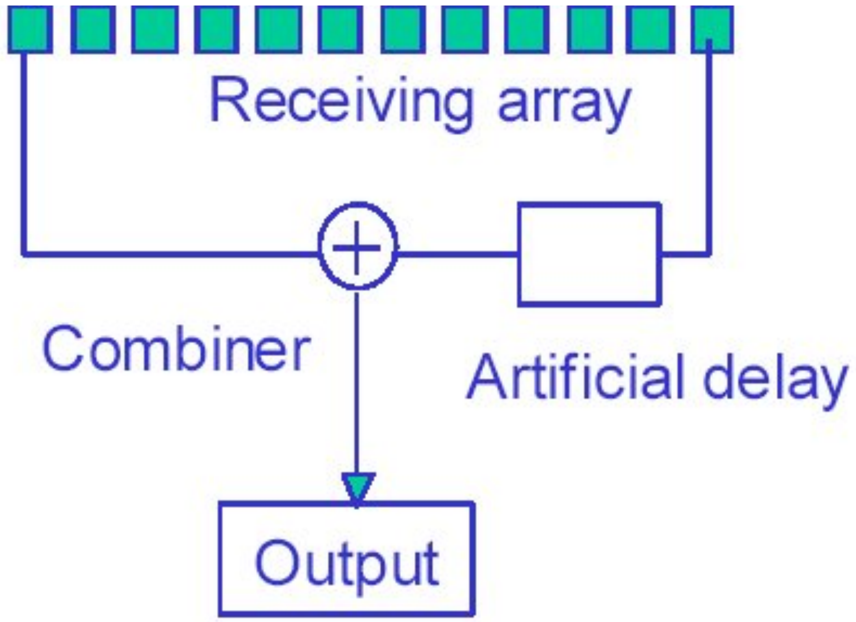
Technology for SKA Low

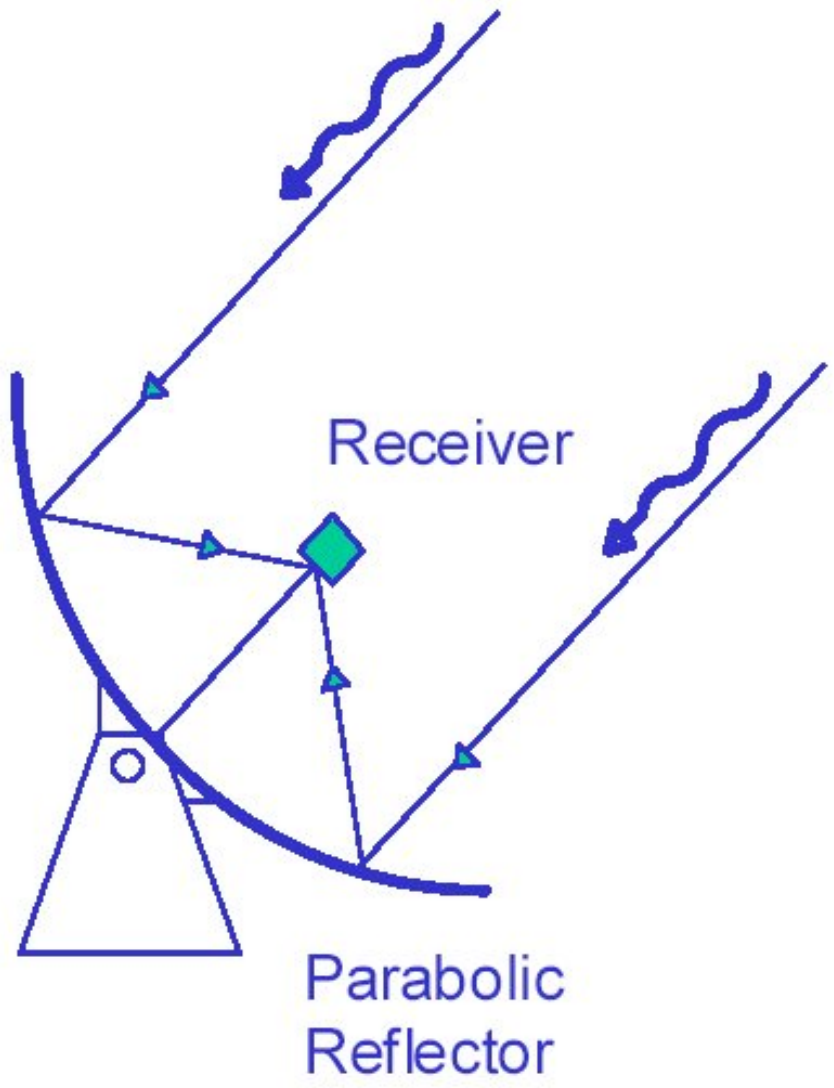





Effelsberg
Nancay
Garching
Chilbolton
Tautenburg
© M. Pluto (2009)
Potsdam

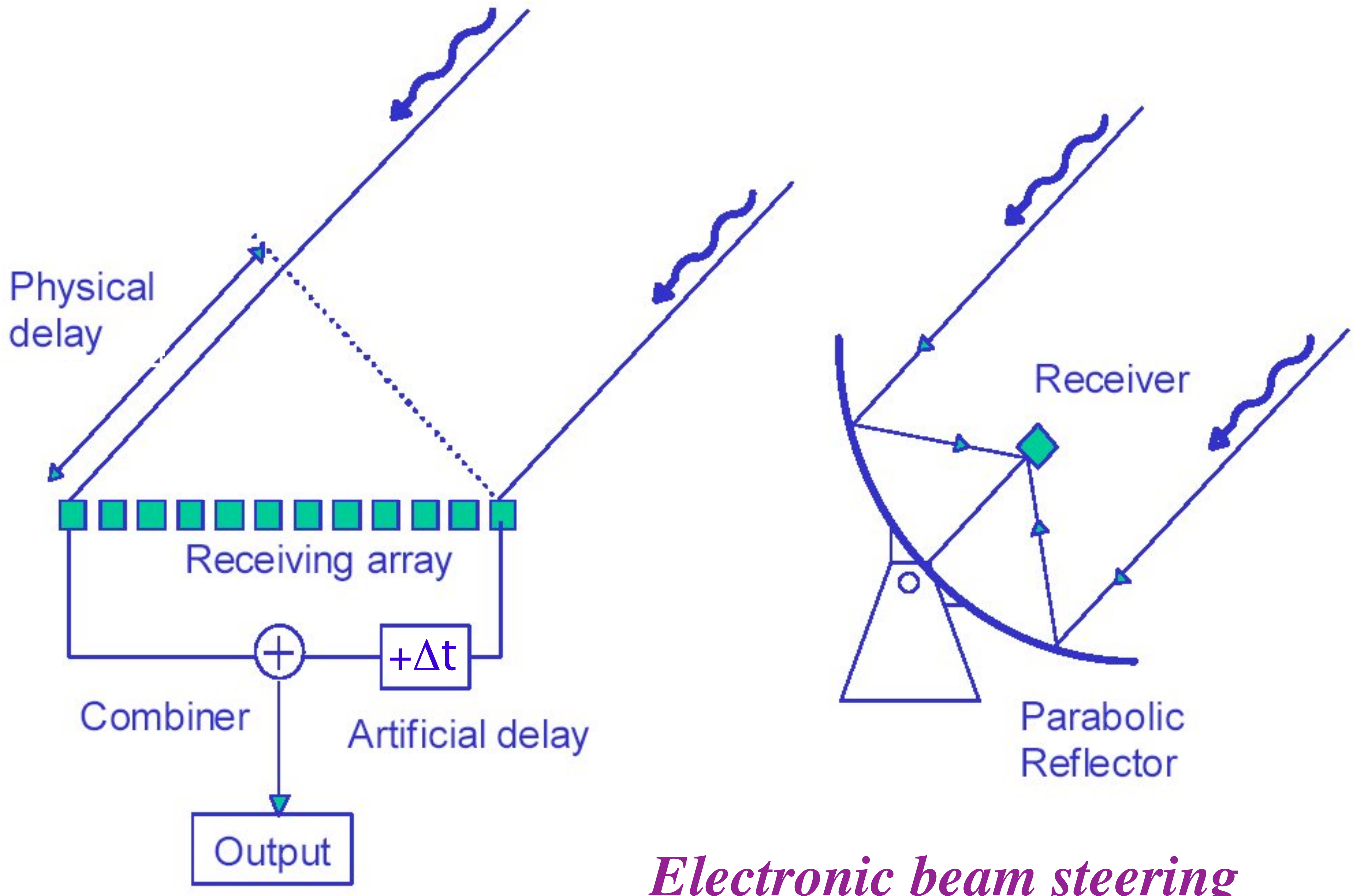
Phased Array Detectors





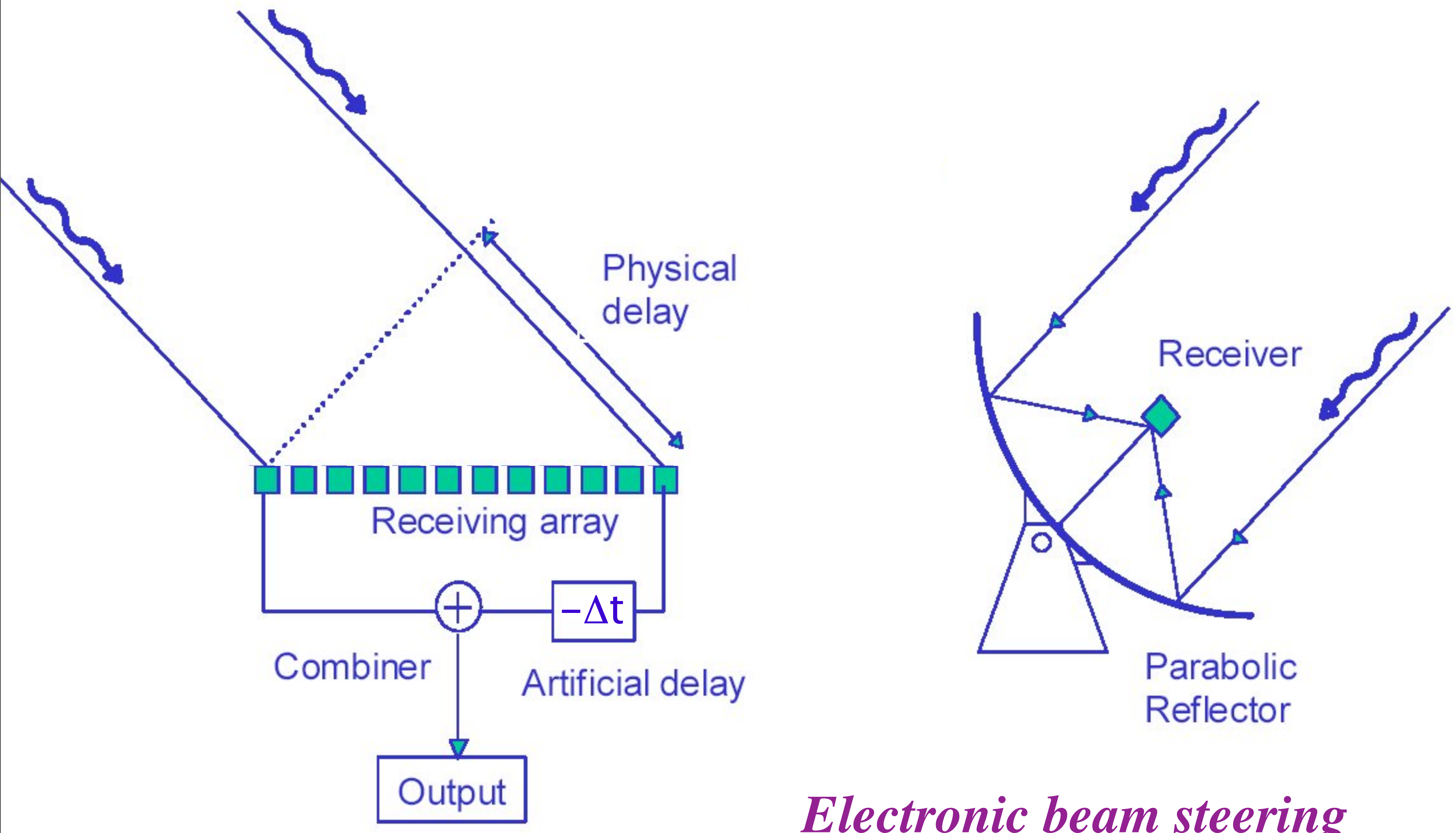
Electronic beam steering

Phased Array Detectors



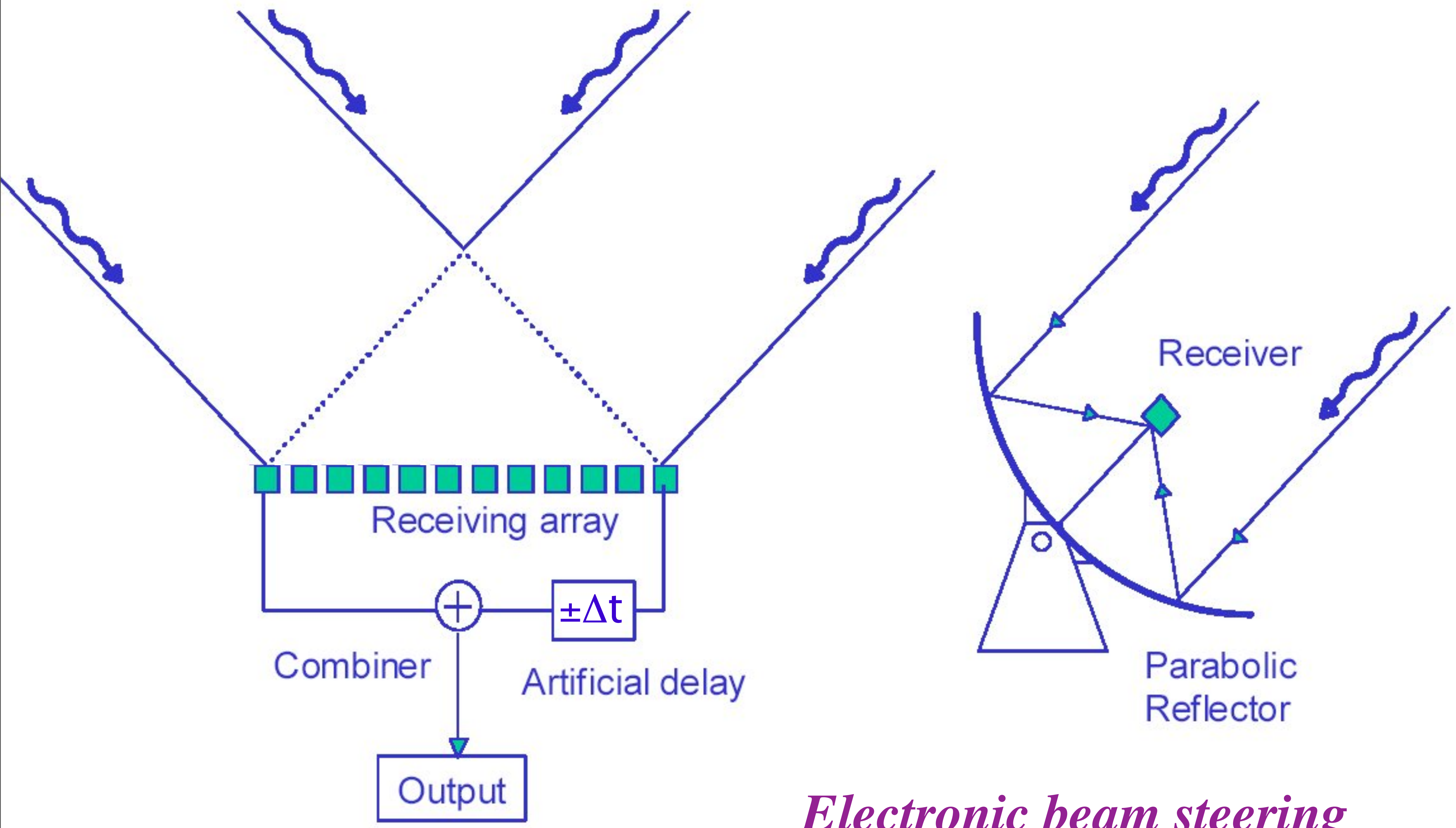
Electronic beam steering

Phased Array Detectors



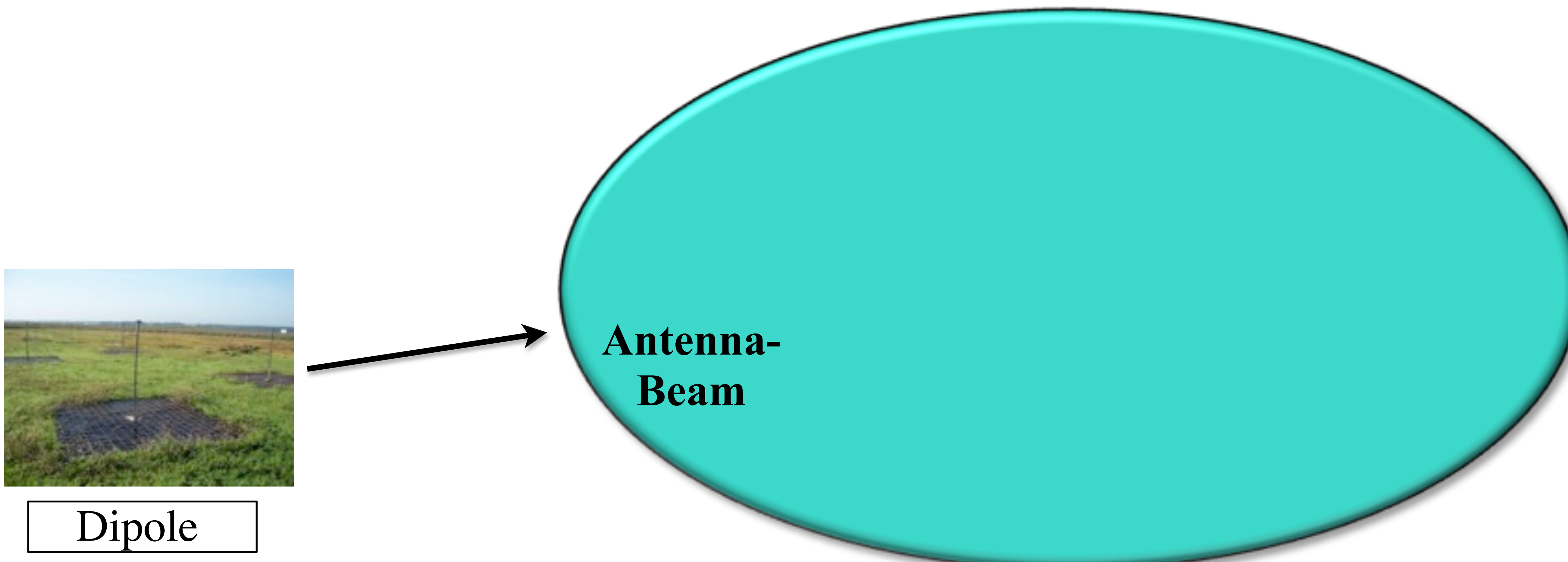
Electronic beam steering

Phased Array Detectors



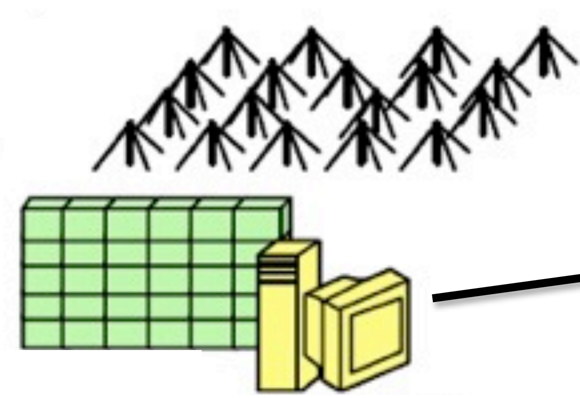
Electronic beam steering

Digital Beam-Forming



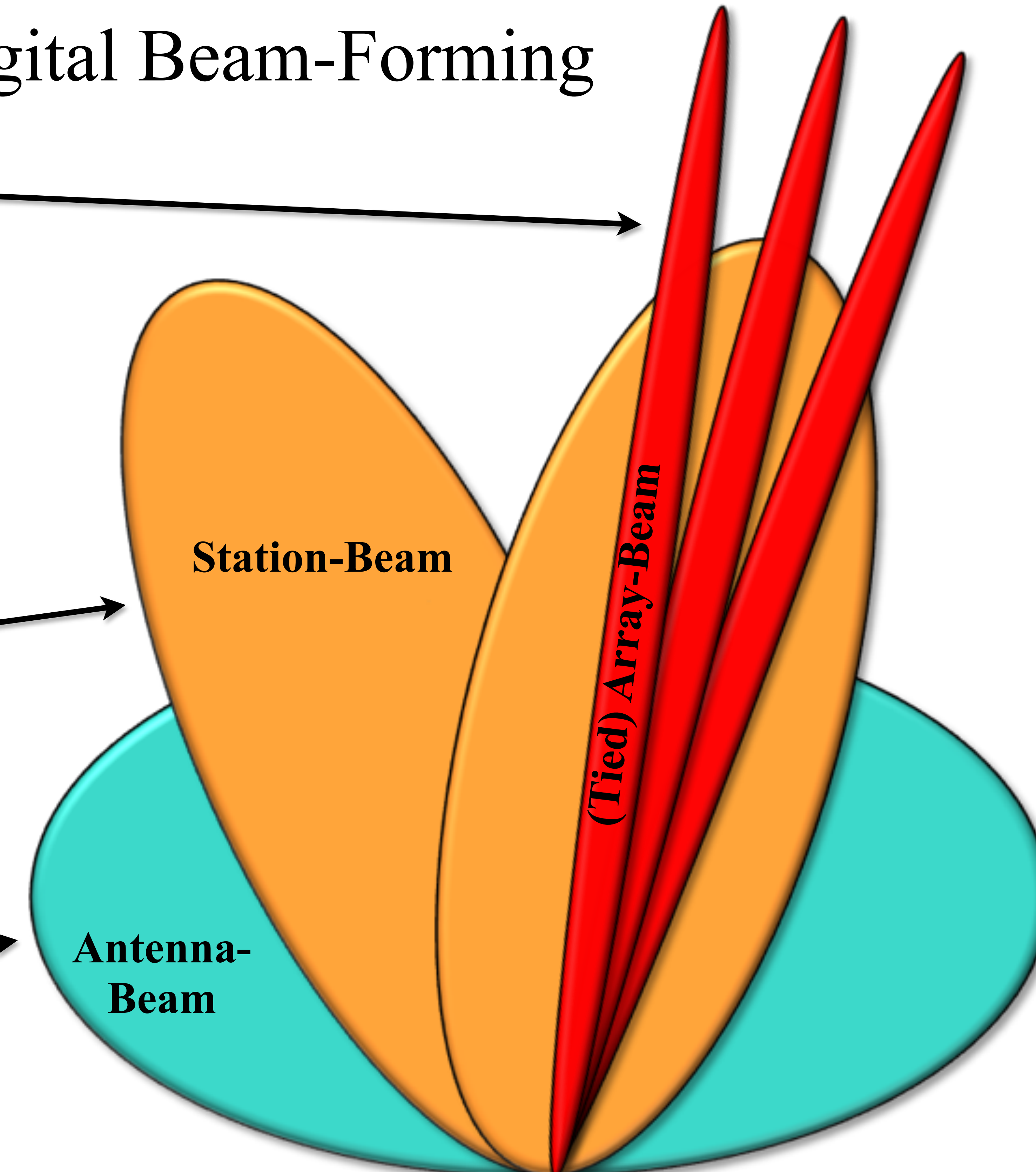
Digital Beam-Forming

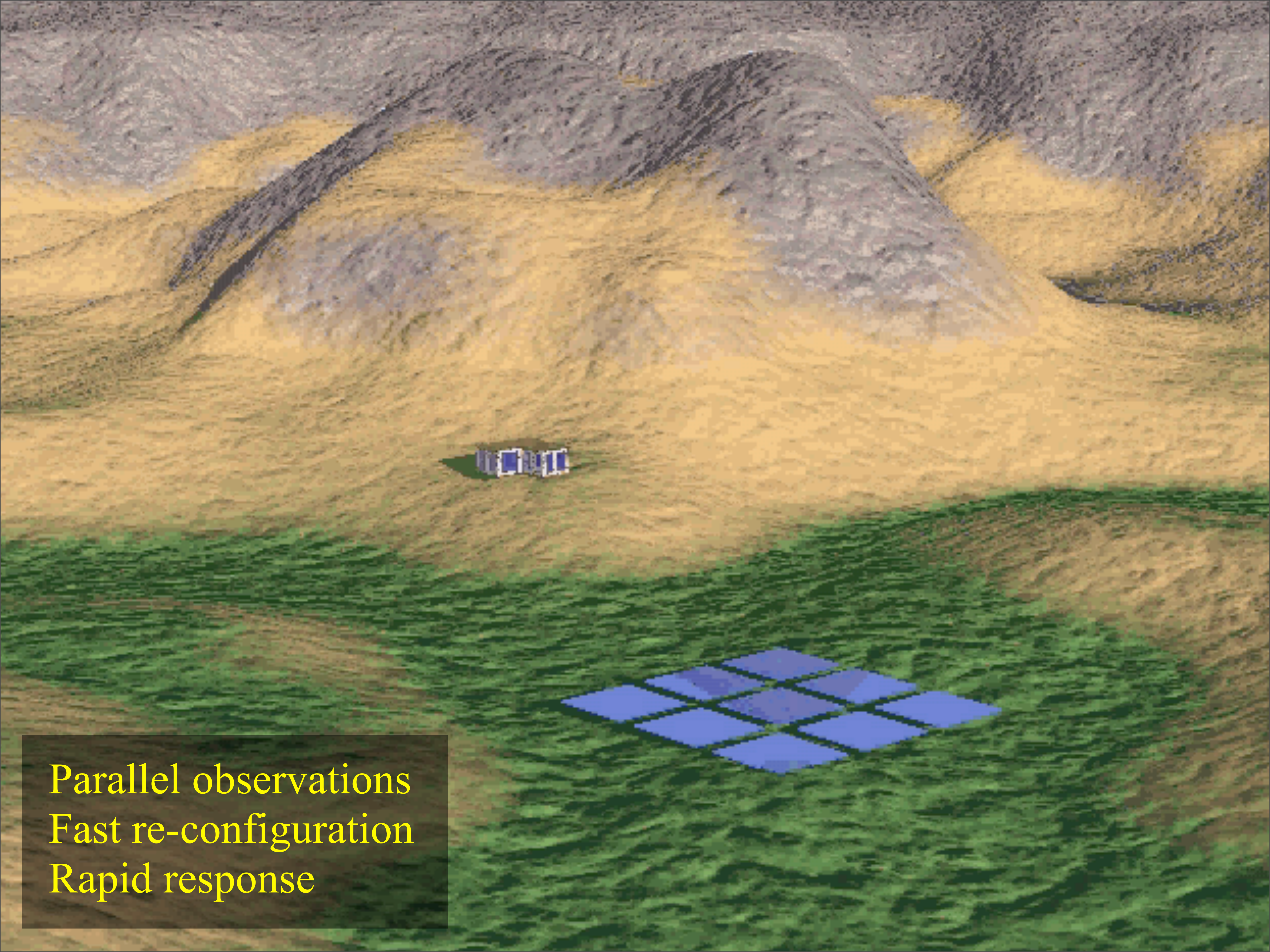
Array



Station

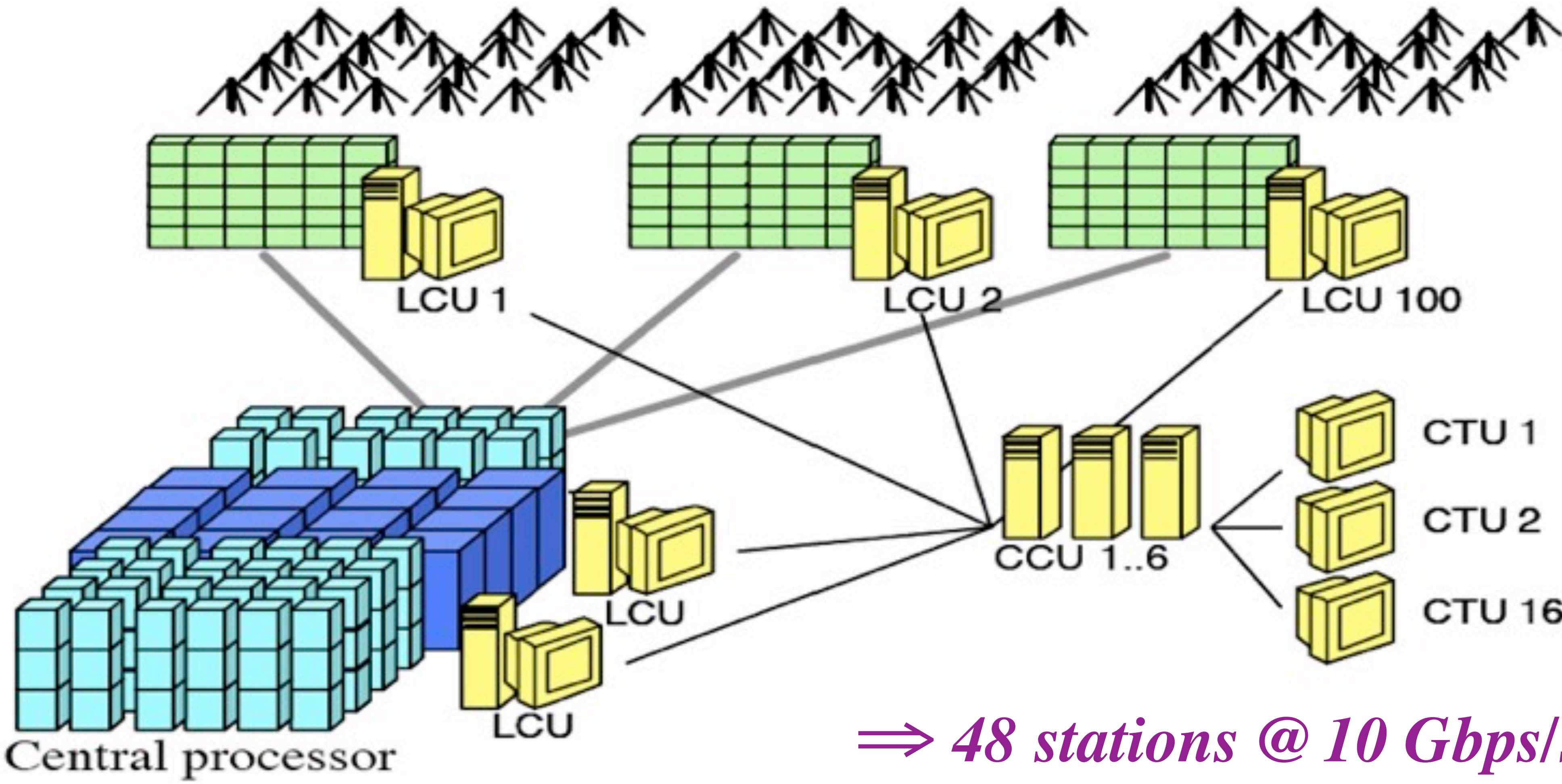
Dipole




Parallel observations
Fast re-configuration
Rapid response

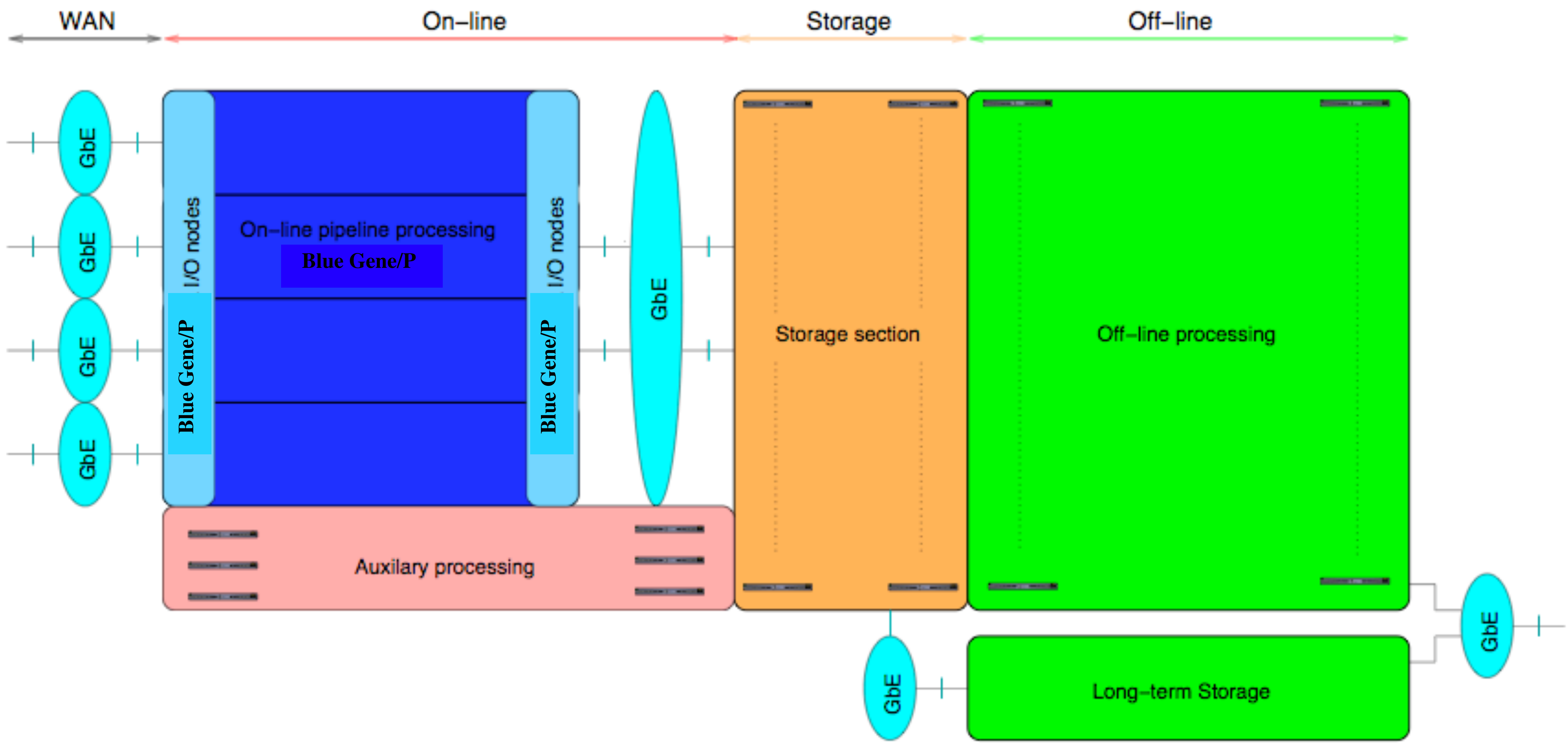
LOFAR Data Flow

- Station level processing *Amplification, digitization, filtering, beam-forming, transient ram buffers (TBB)*
    - Central processing *Delay compensation, correlation, calibration, science pipelines (BG/P, storage, offline cluster)*



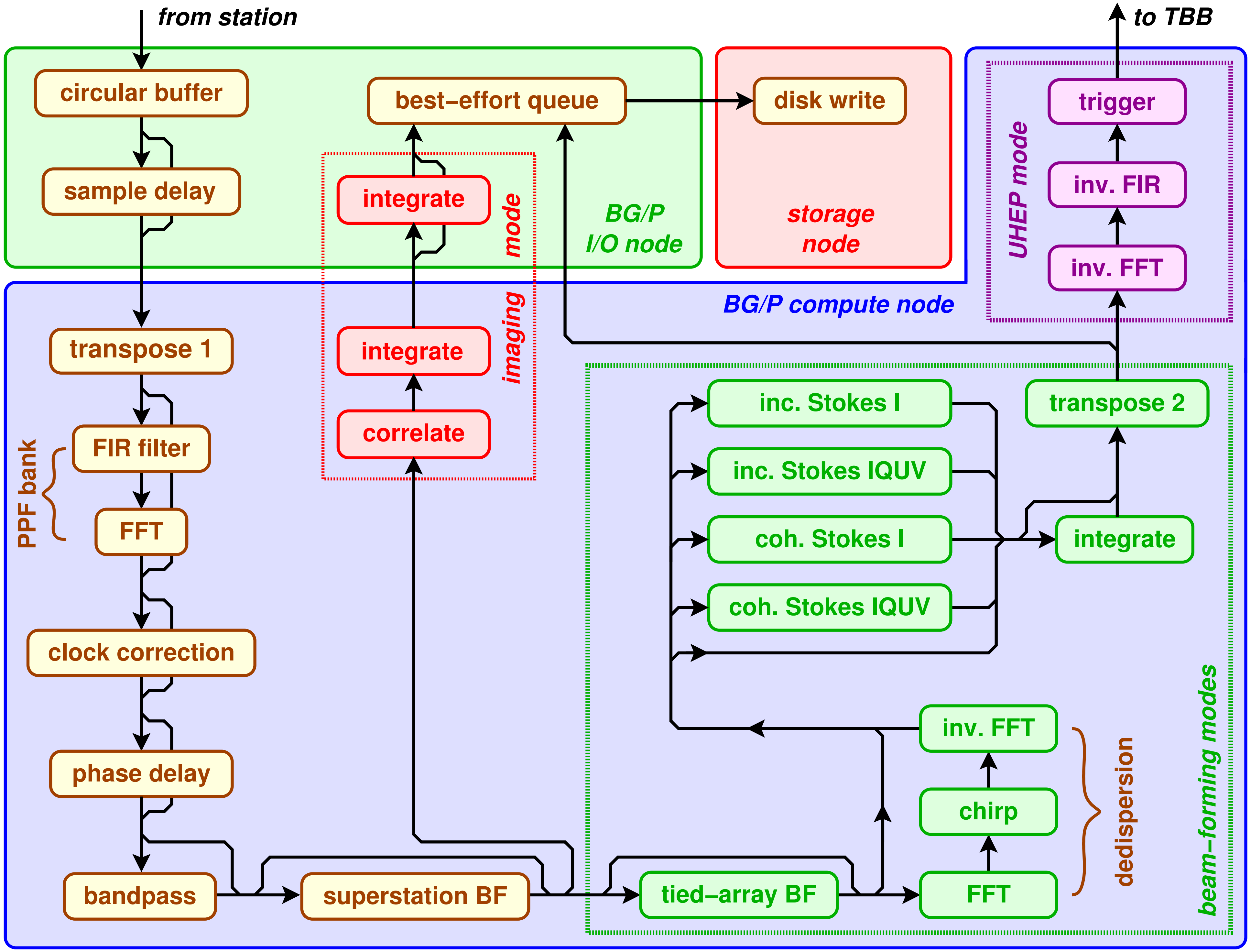
⇒ 48 stations @ 10 Gbps/station

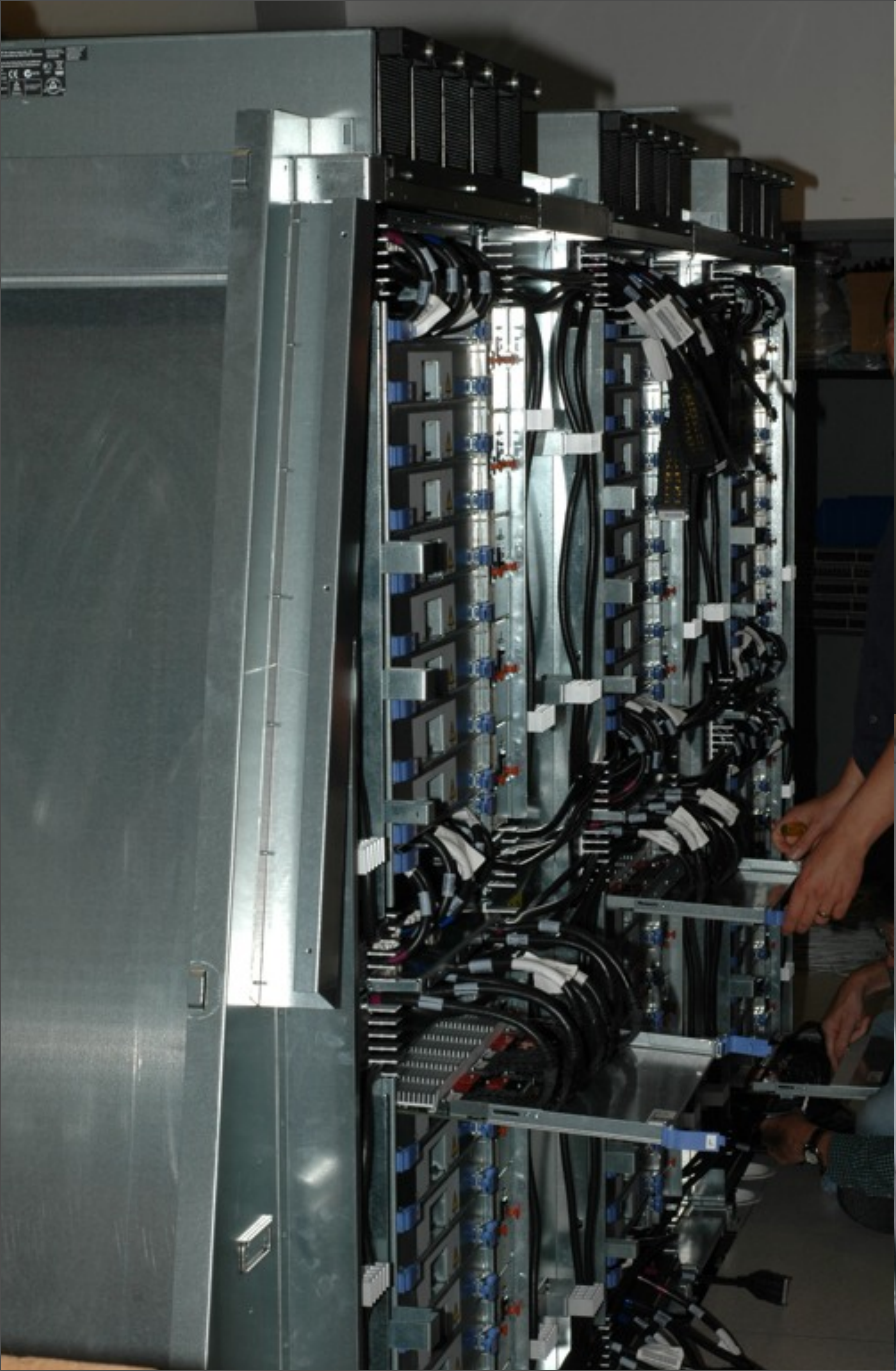
Central Processing



- BG/P *Data reception, transpose, correlation, beam-forming, de-dispersion, 45 TFLOPS*
  - Storage system *Short term storage of data, ~2 PByte, ~100Gbps I/O*
    - Offline cluster *Pipelines, data products, off-line analysis, ~20 TFLOPS*

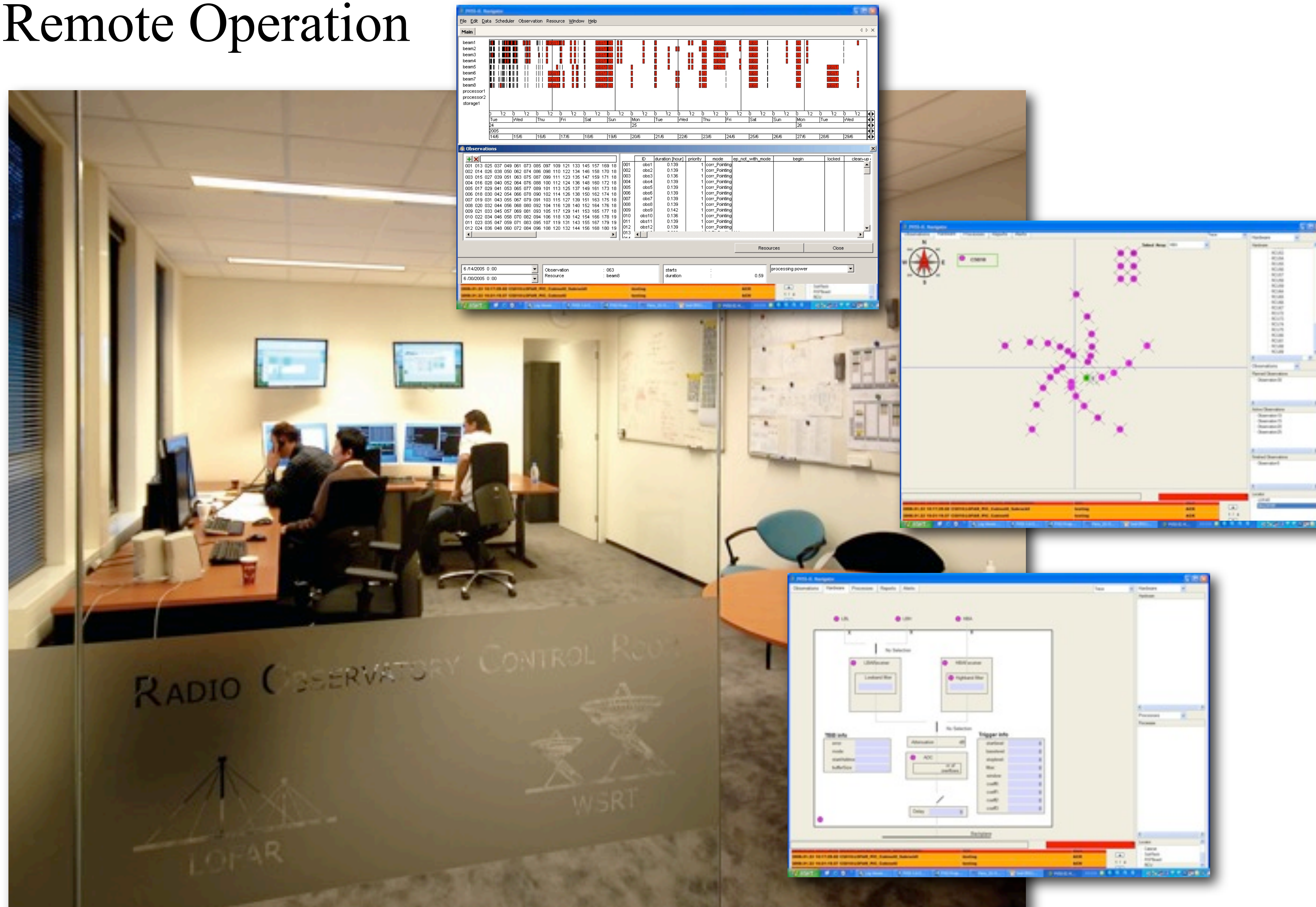
Online Processing






IBM
Blue Gene/P

Remote Operation



Some numbers

- 2688 dipoles (LBA), 200 MHz sampling, 2 polarizations, 12 bit digitization
  ⇒ 13 Tbits/s ~ 1.6 TB/s ~ 138 PB/day

- 48 stations, 48 MHz total bandwidth, 8 independent beams (up to 100s)

- 1128 baselines, 242 sub-bands, 256 channels, 4 polarizations, 1 sec correlator dump-time ⇒ ~ 10 TB/hr ~ 240 TB/day ~ 0.1 EB/yr

Storage limits give a ~1 week processing window

LOFAR Science Drivers

Key Science Projects

Epoch of Reionization

Transients and Pulsars

High Energy Cosmic Rays

Surveys and the Distant Universe

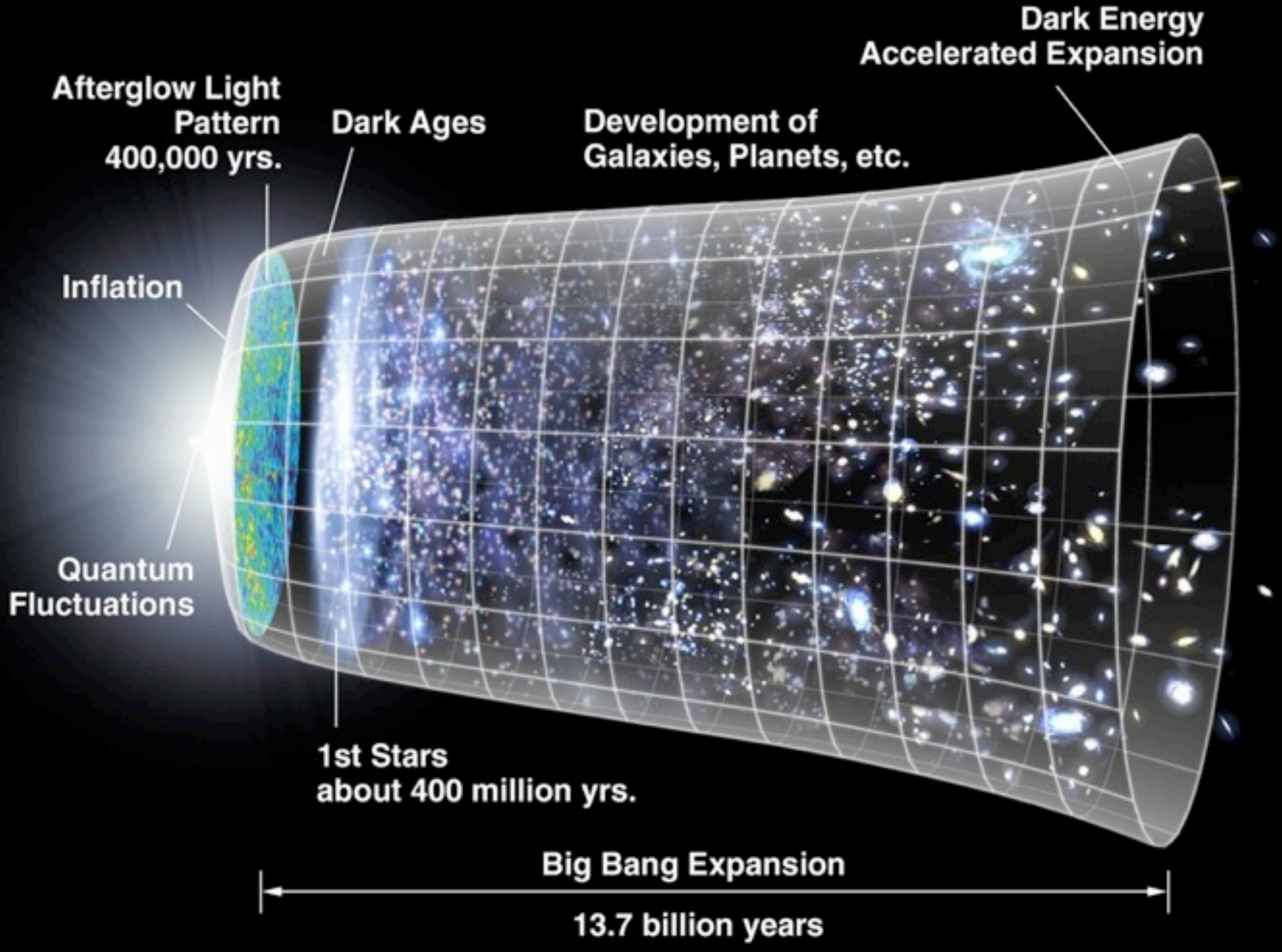
Cosmic Magnetism

Solar Physics and Space Weather

⇒ International membership from countries all over world

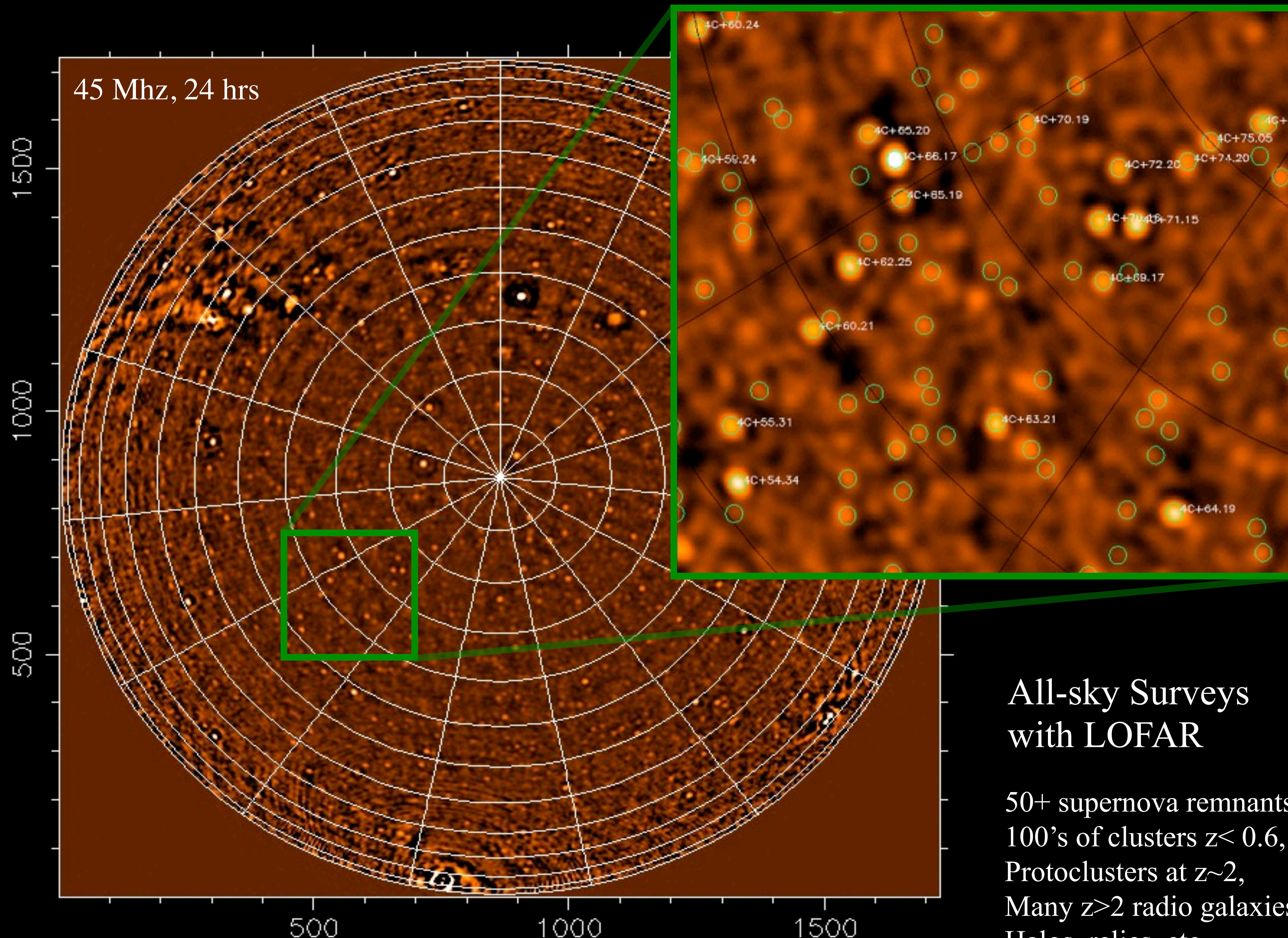
Contribute development and commissioning resources

The LOFAR Epoch of Reionization Key Science Project



Goal: Tracing the EoR in HI and possibly the late stages of the Dark Ages

$\Rightarrow$ 1.5 Pbytes and 10^{21}-10^{22} FLOP to extract signal!

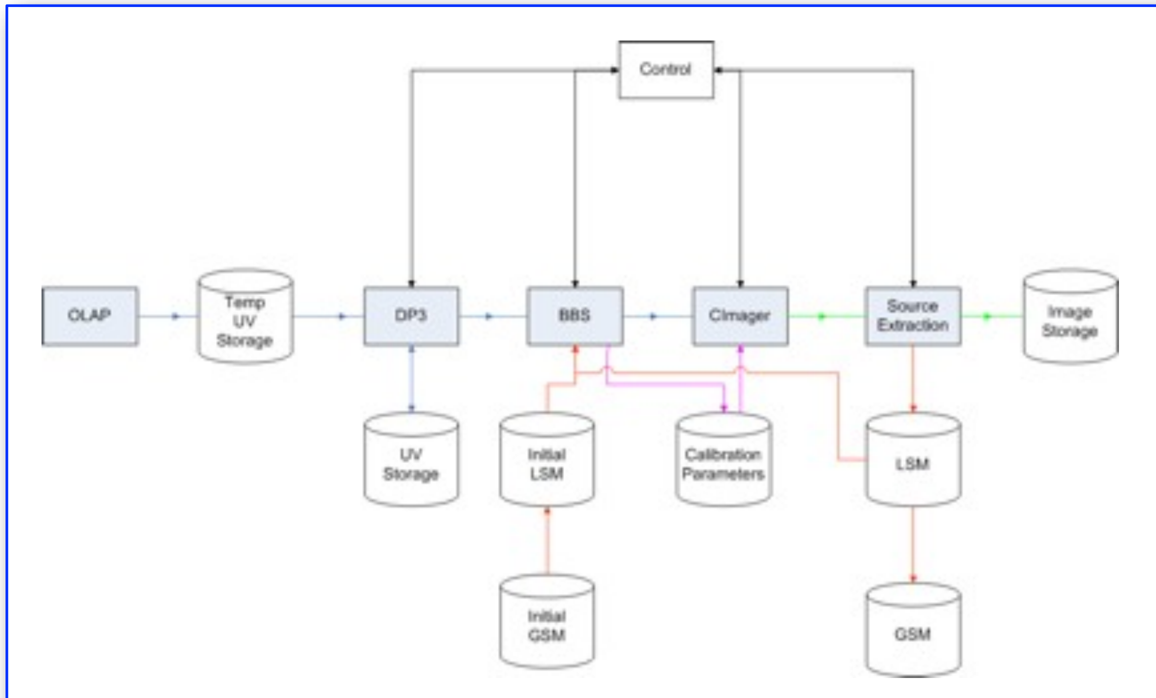


All-sky Surveys with LOFAR

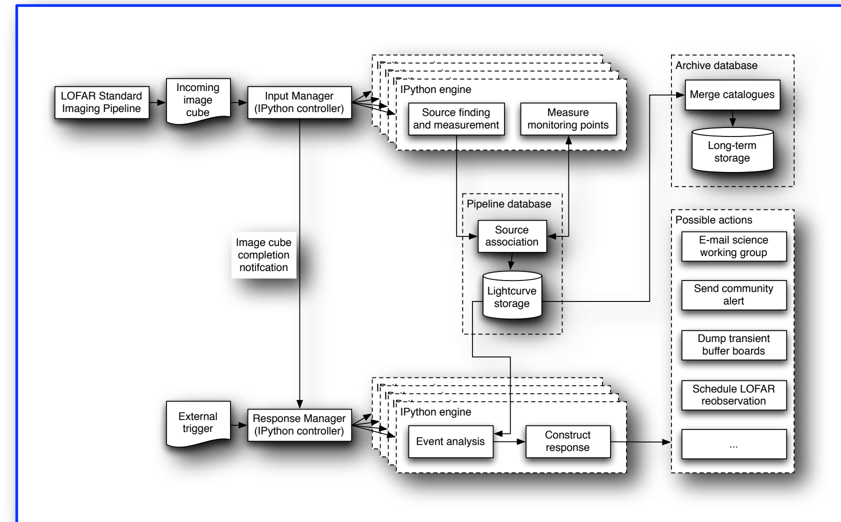
50+ supernova remnants,
100's of clusters z< 0.6,
Protoclusters at z~2,
Many z>2 radio galaxies,
Halos, relics, etc...

Science Pipelines

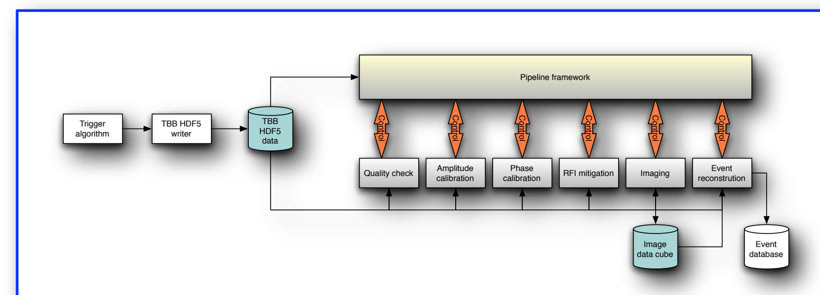
Standard Imaging



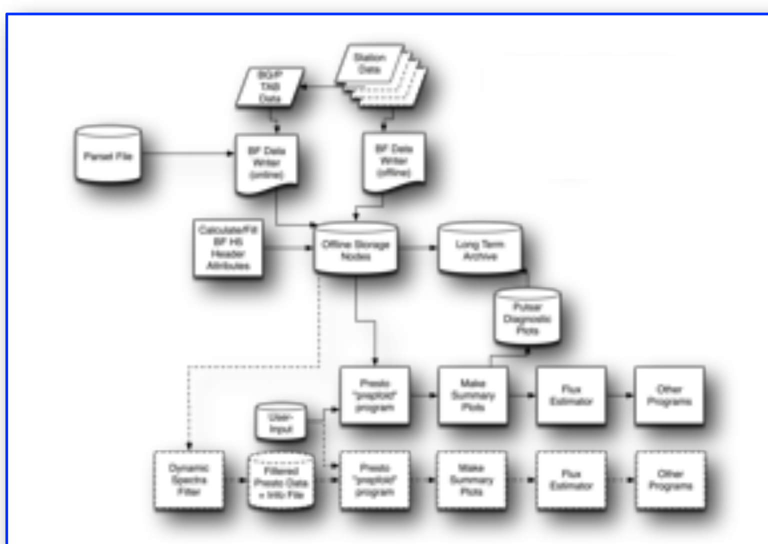
Transient Detection



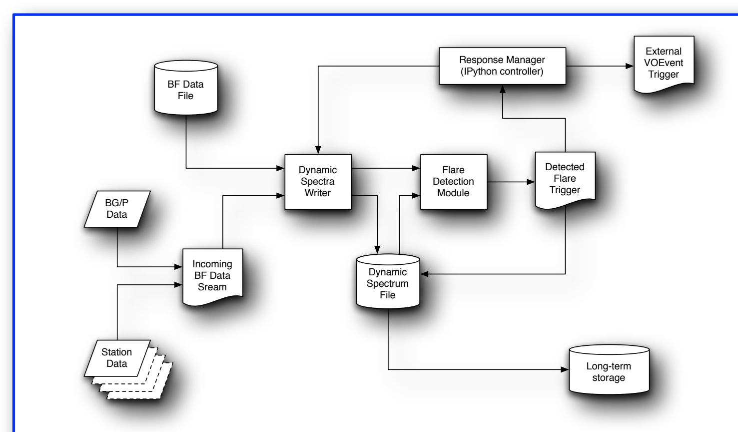
VHECR



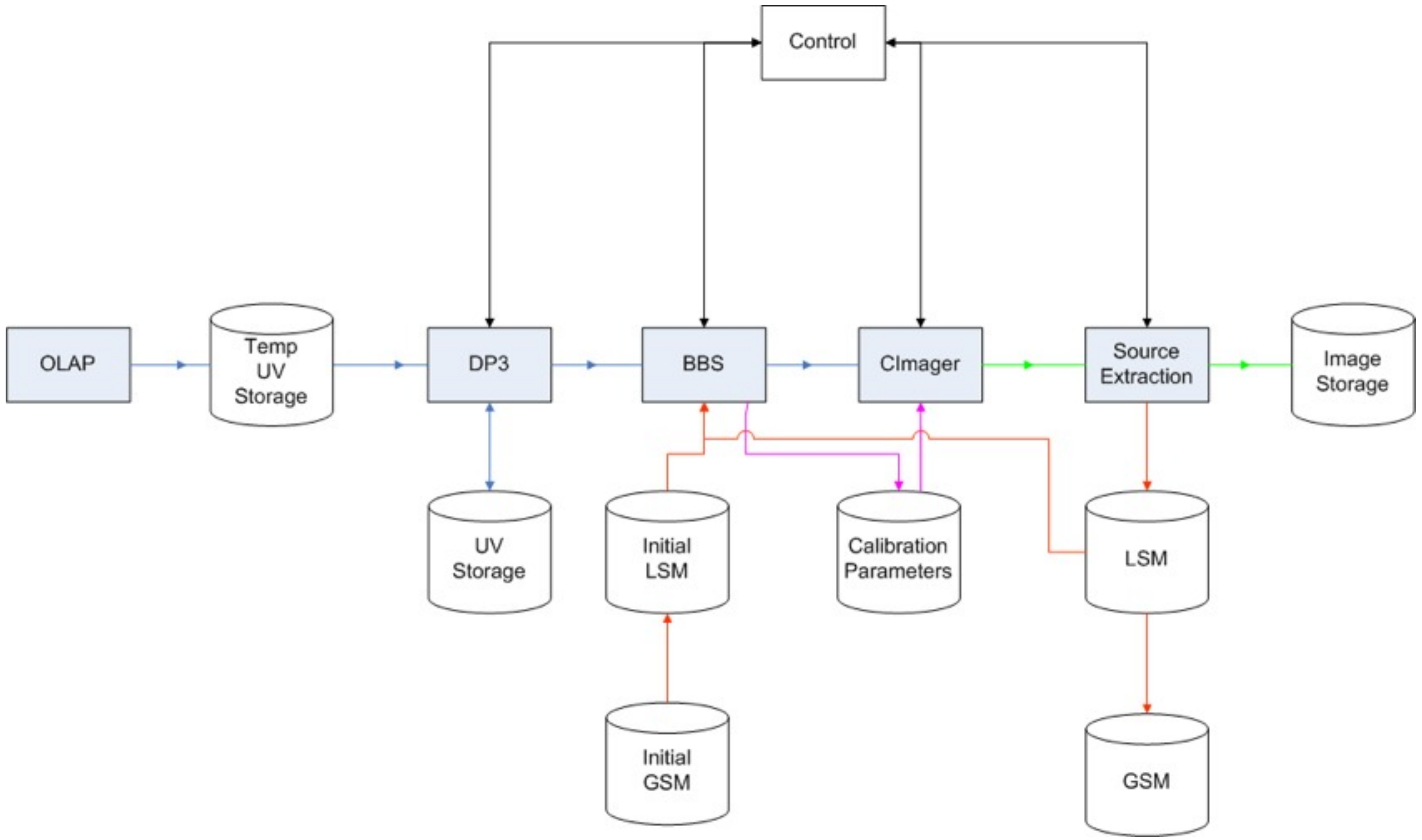
Known Pulsars



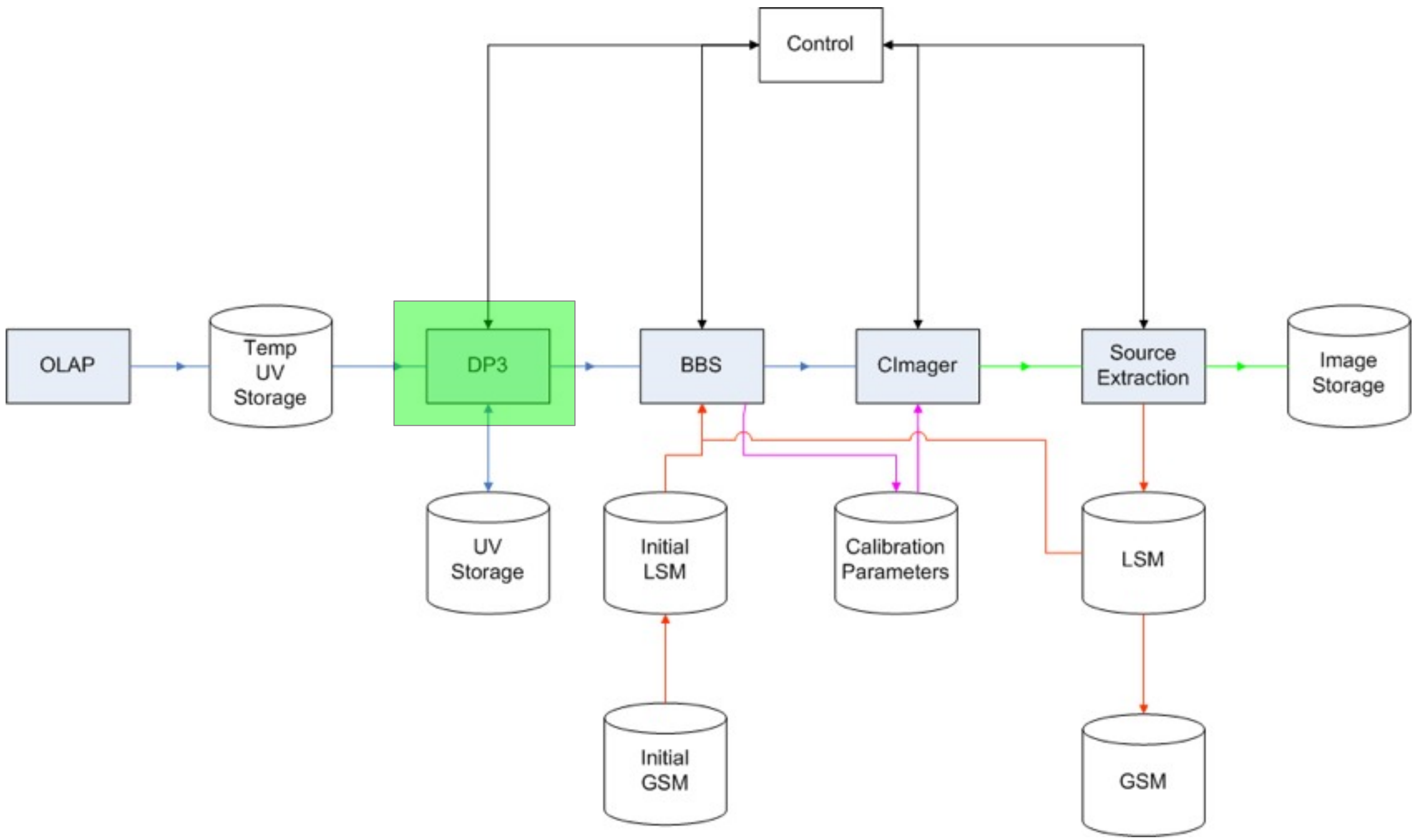
Dynamic Spectra



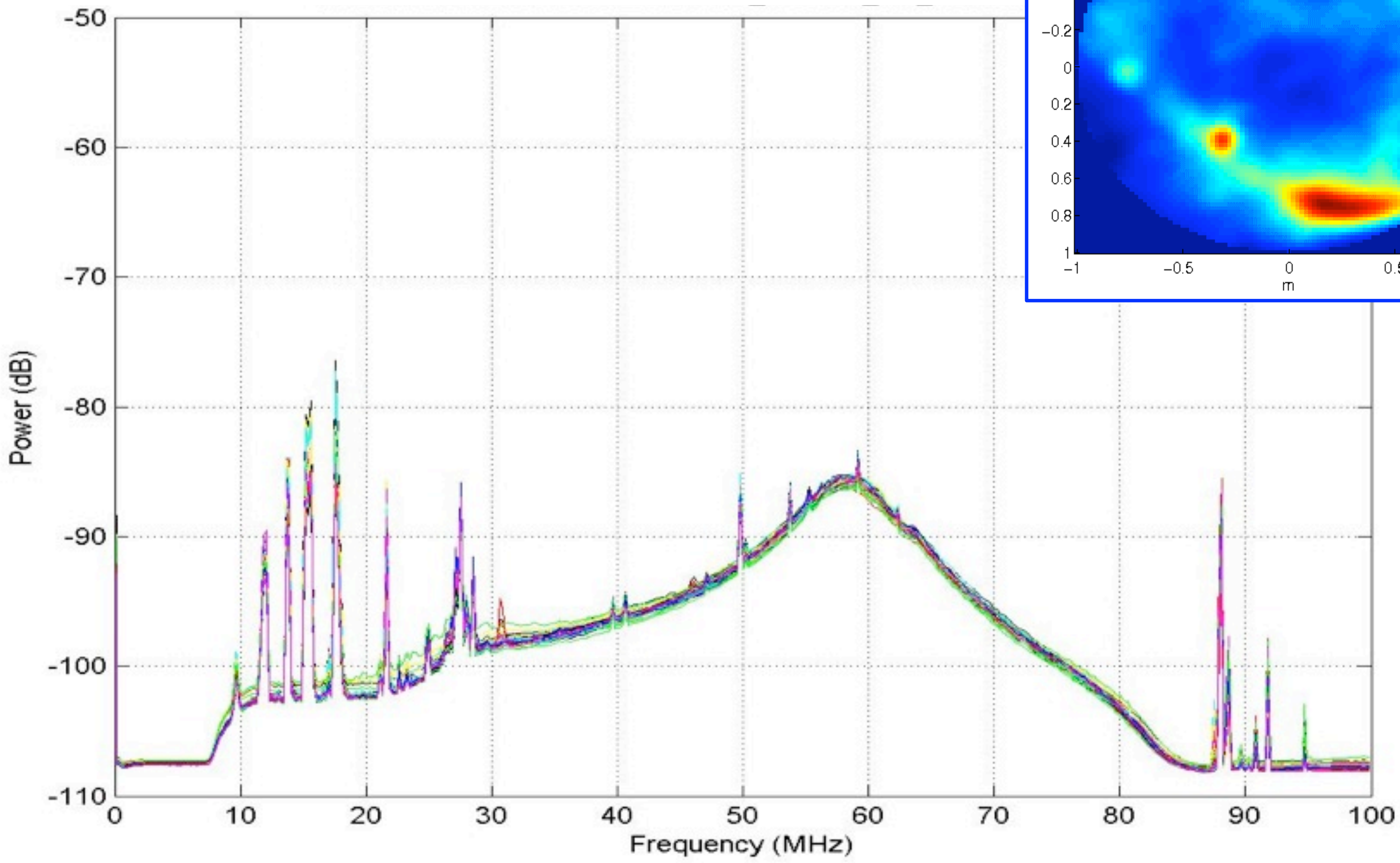
Standard Imaging Pipeline

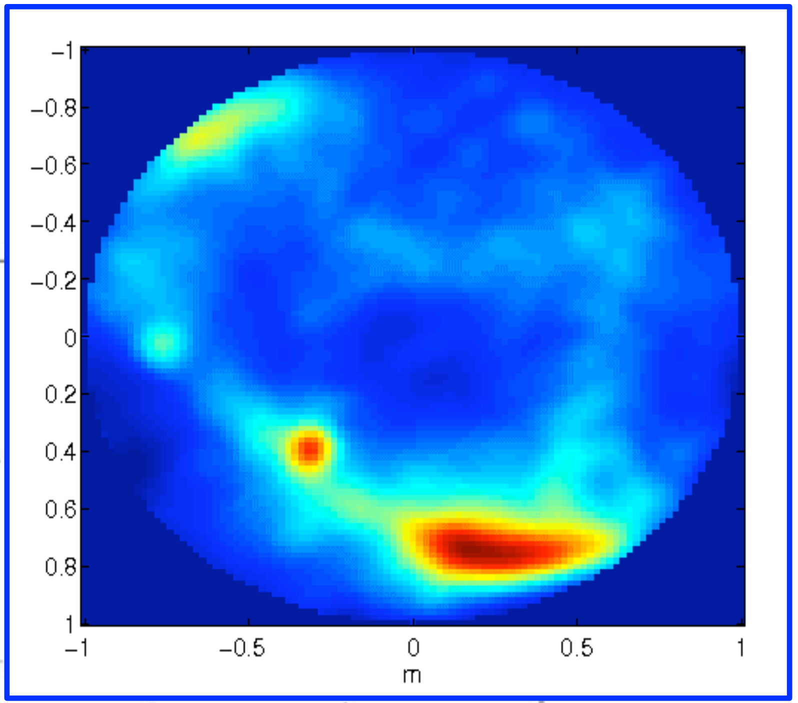


Standard Imaging Pipeline

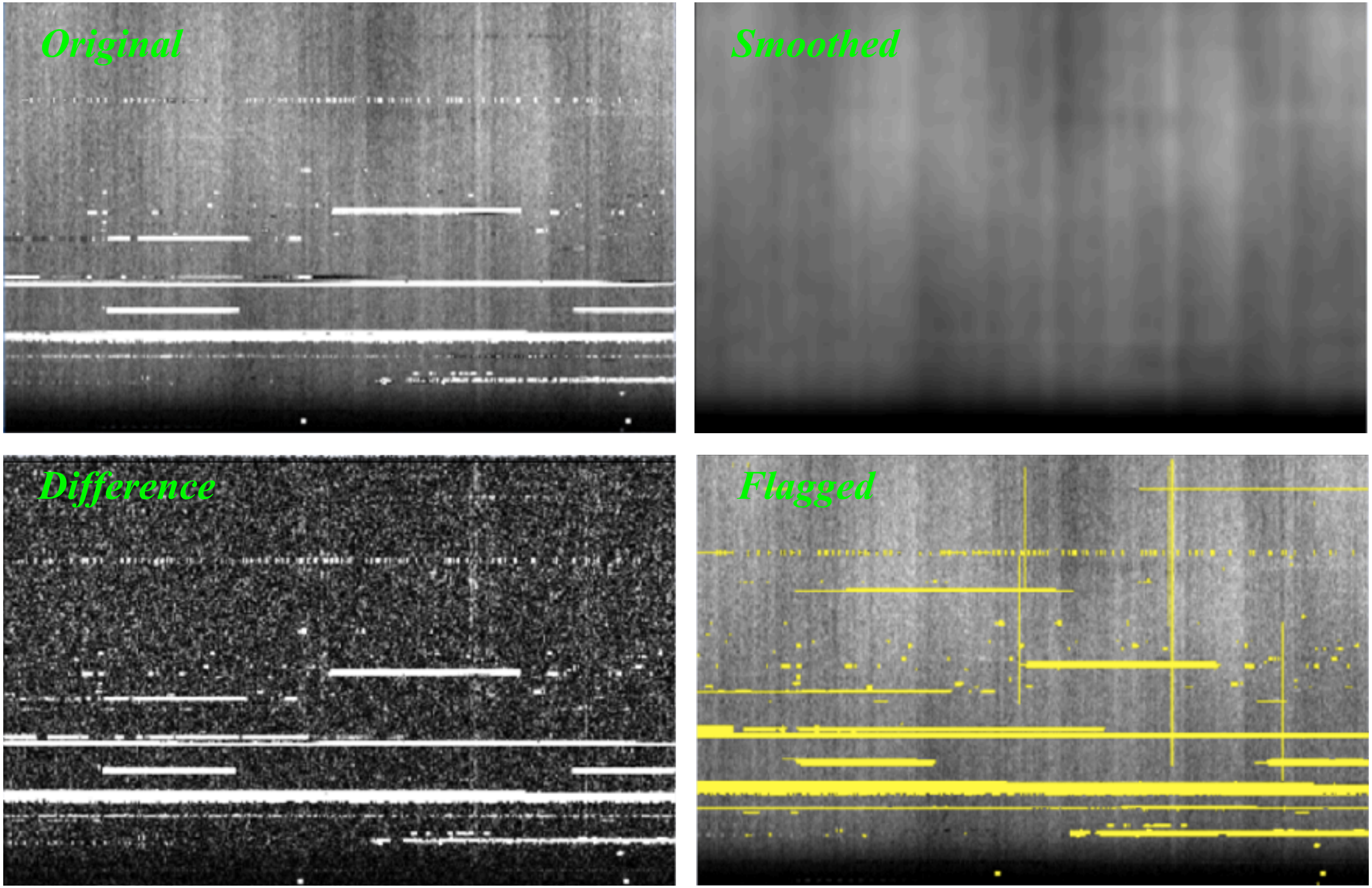


Local RFI Environment

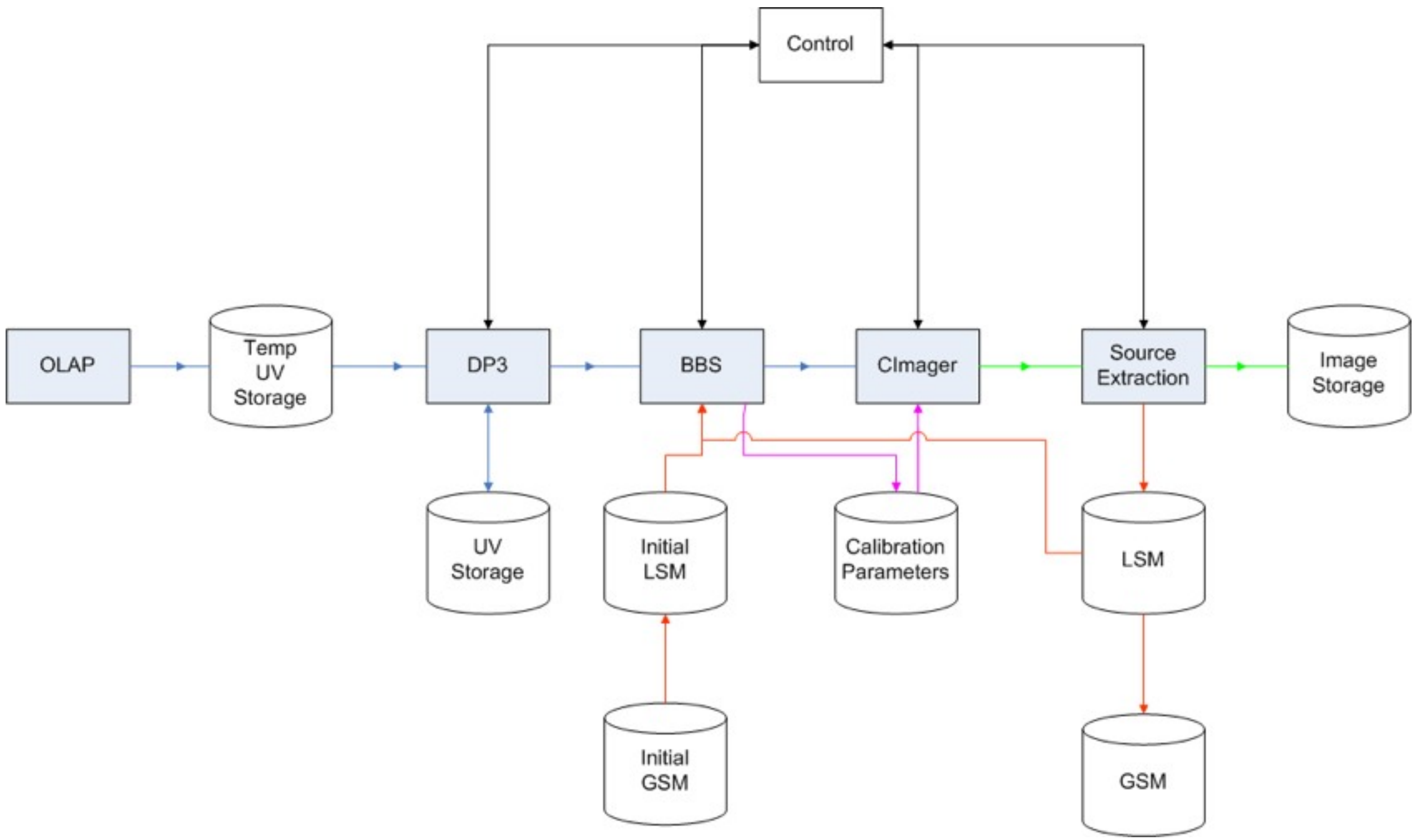




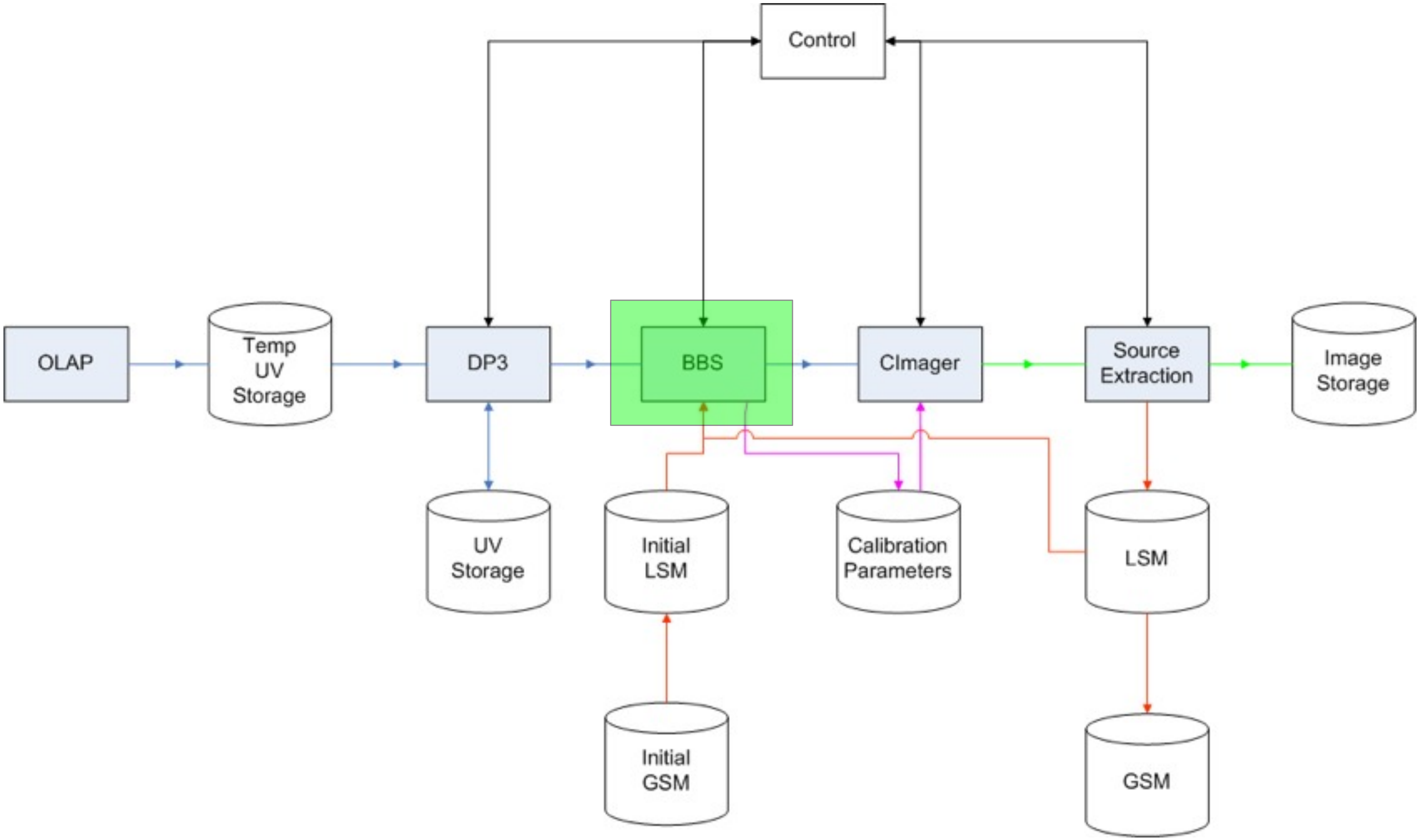
Automated RFI Flagging



Standard Imaging Pipeline

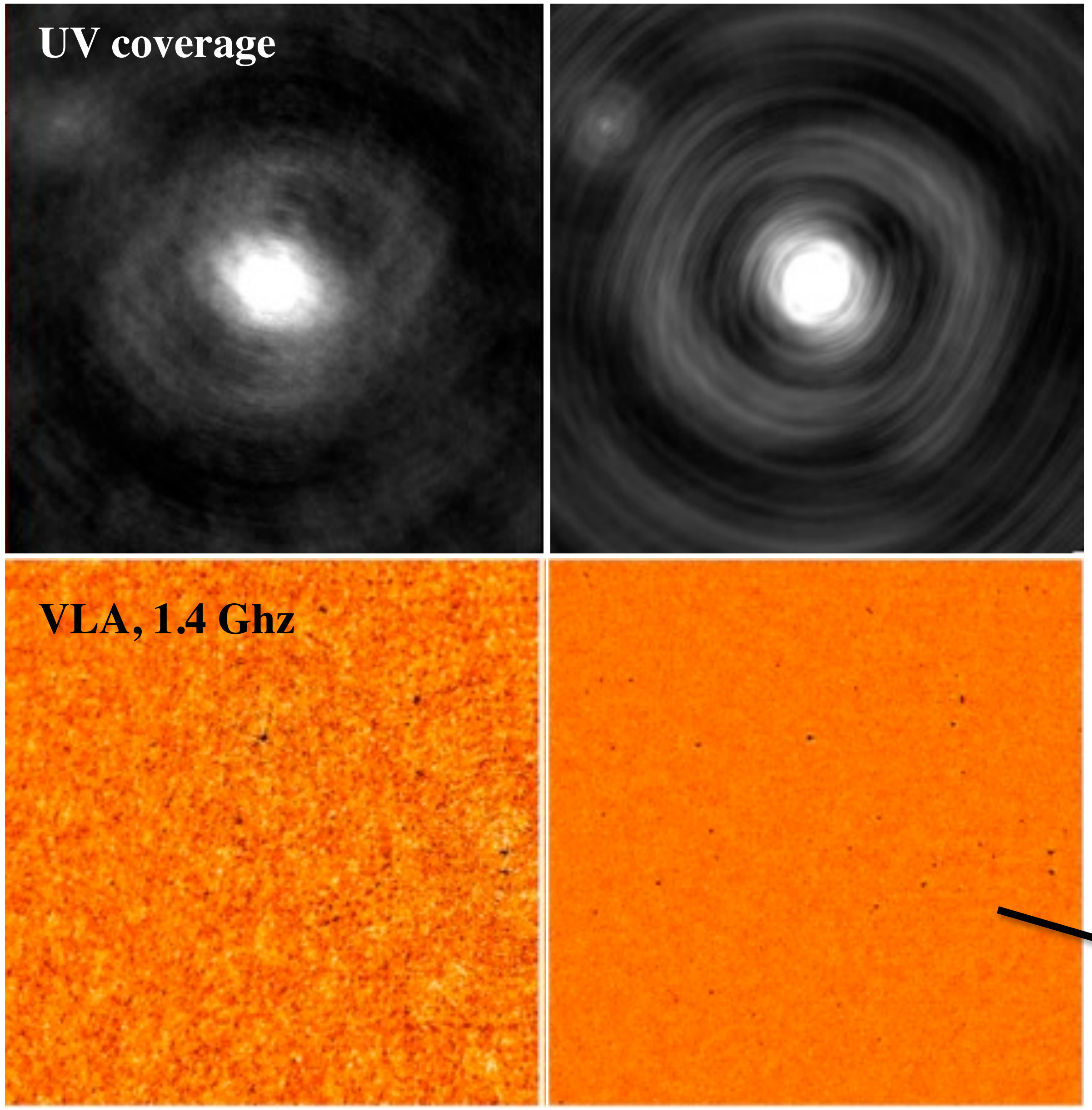


Standard Imaging Pipeline

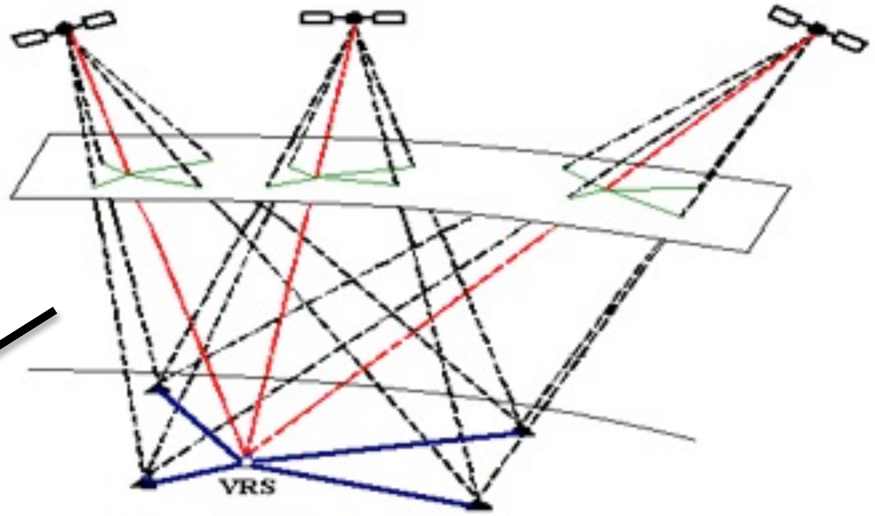


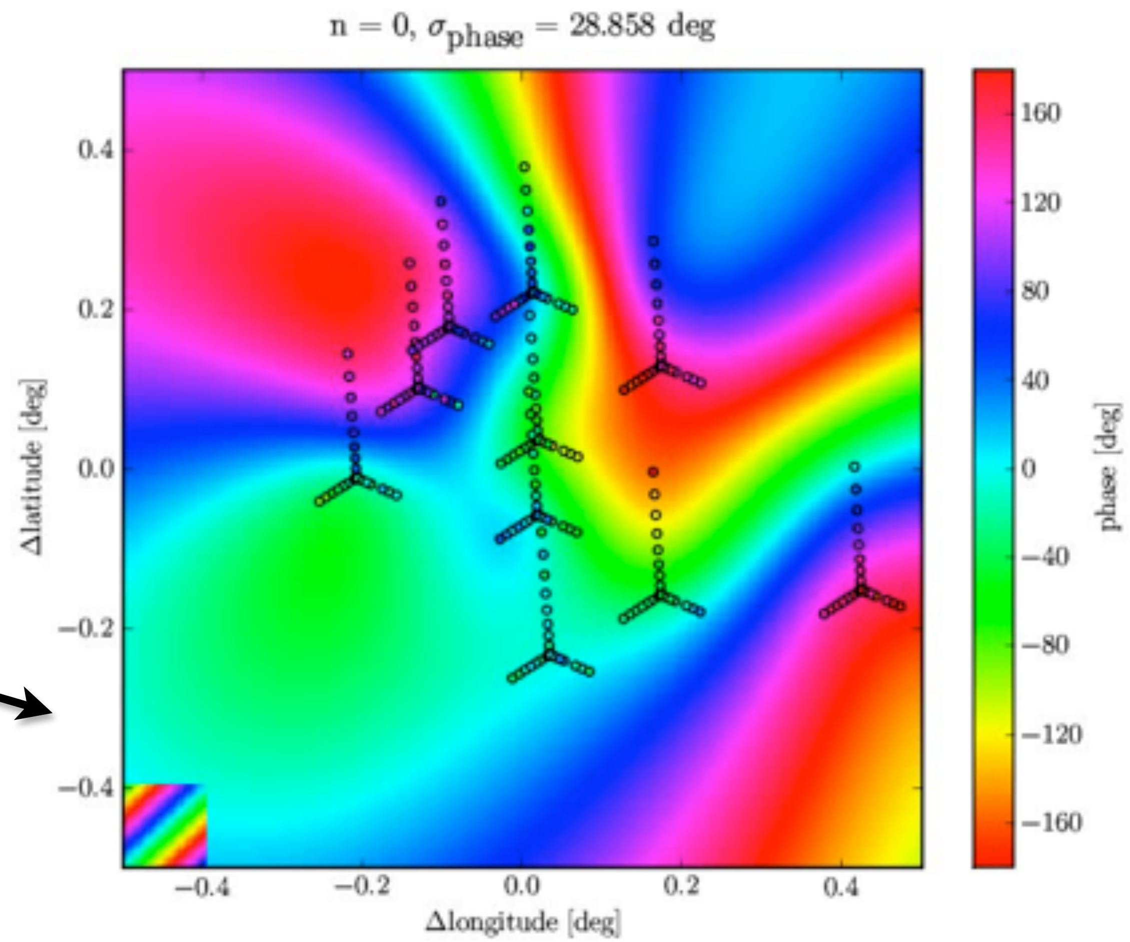
Ionosphere Correction

Uncorrected | Corrected

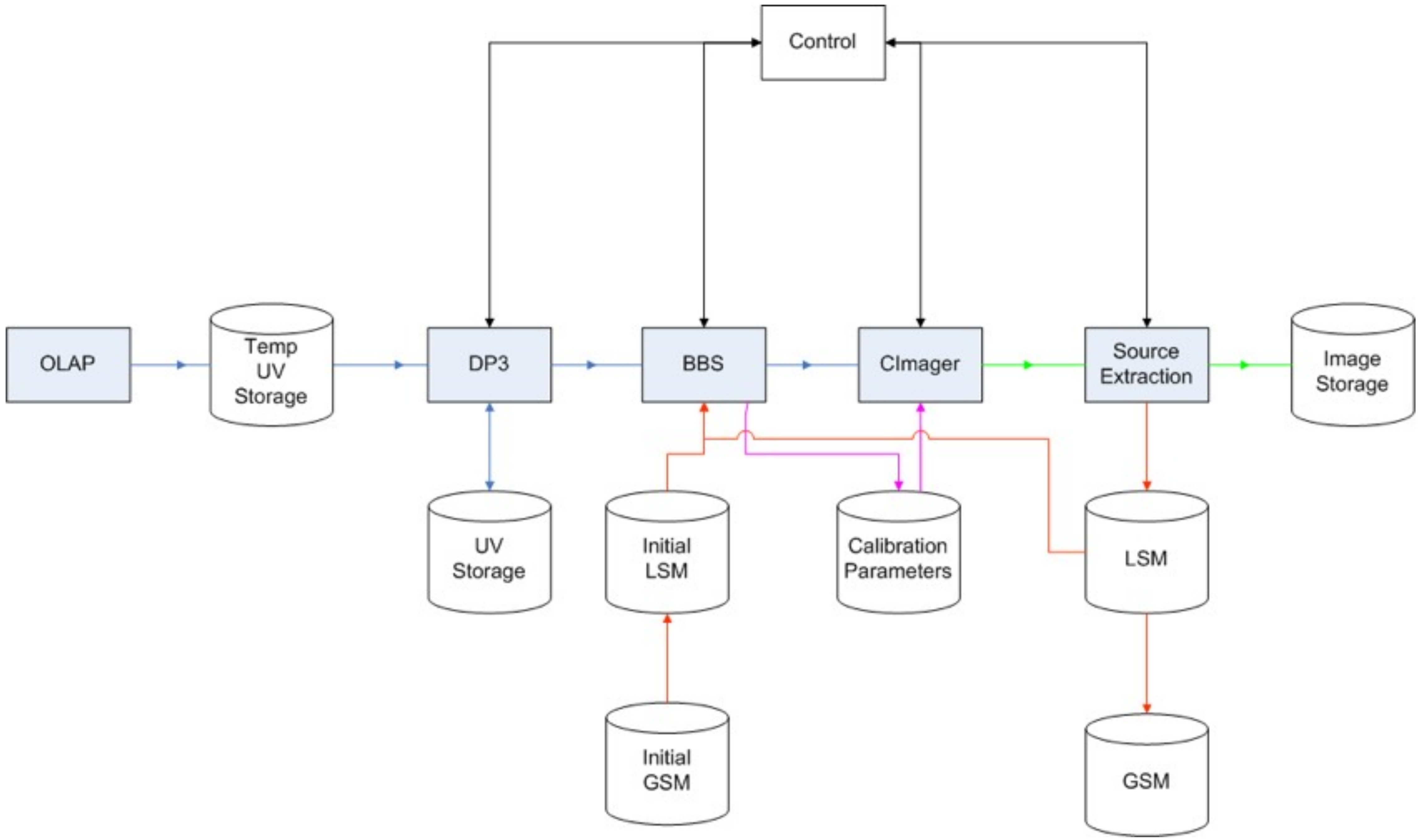


(courtesy H. Intema)

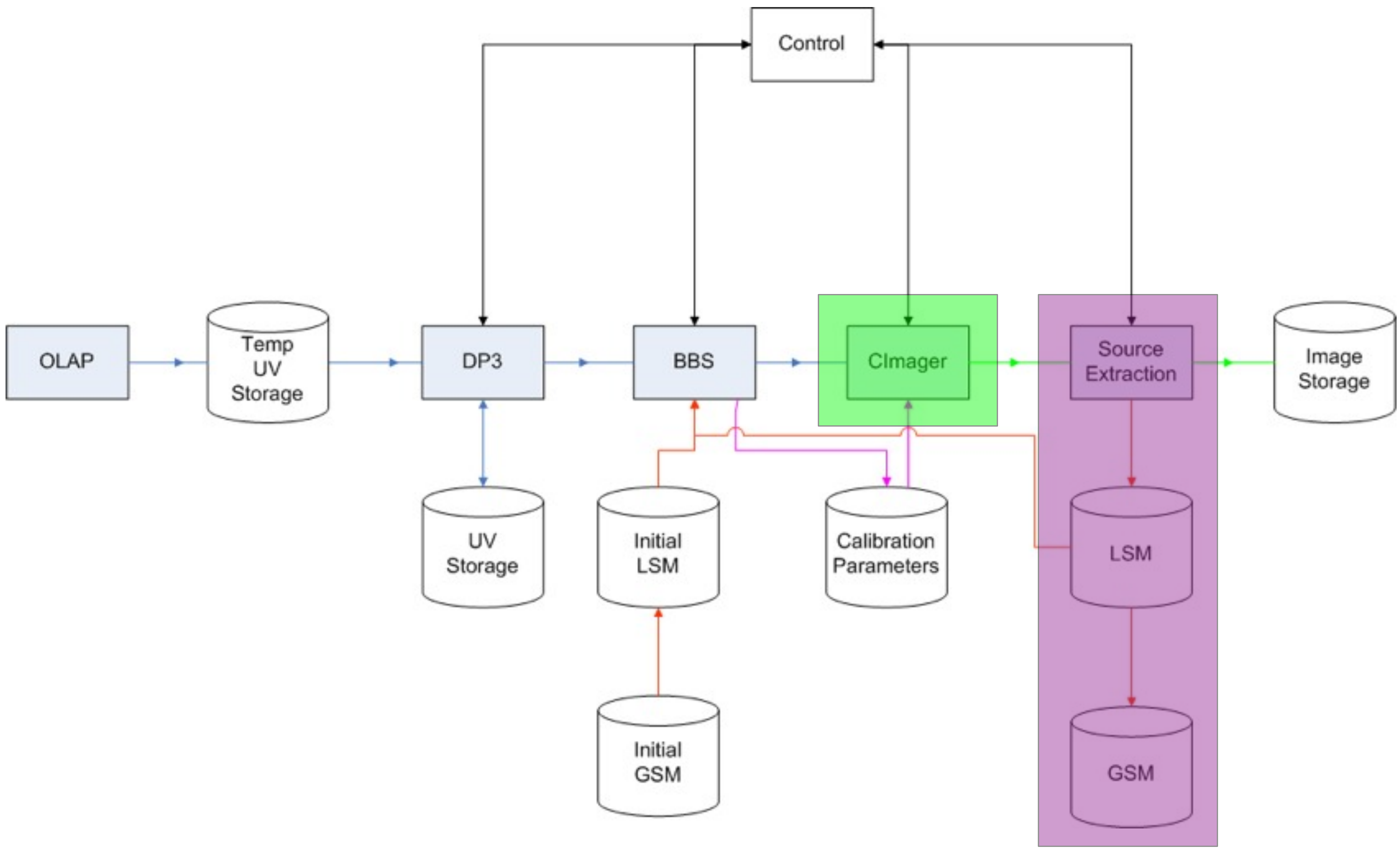


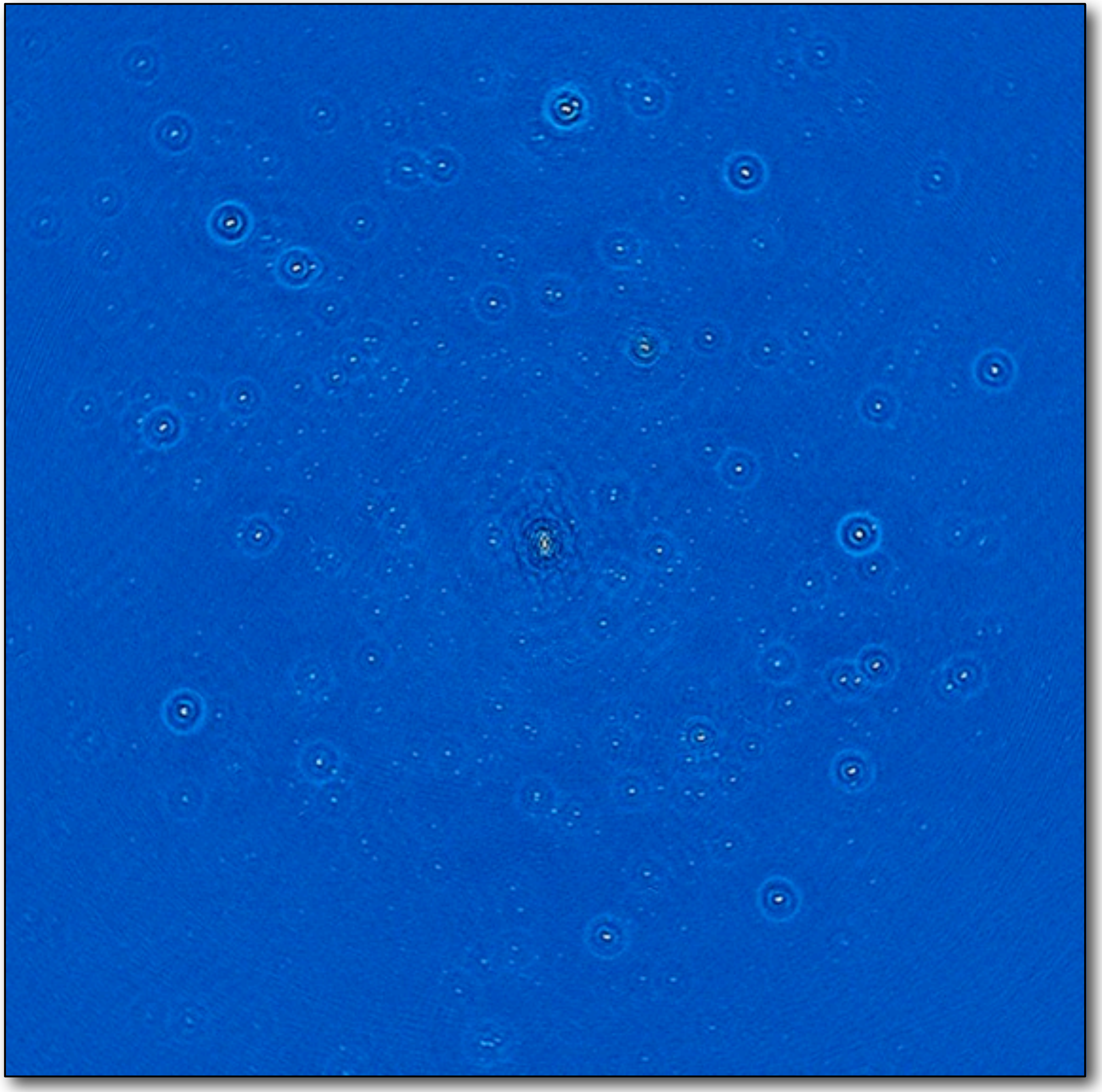


Standard Imaging Pipeline



Standard Imaging Pipeline





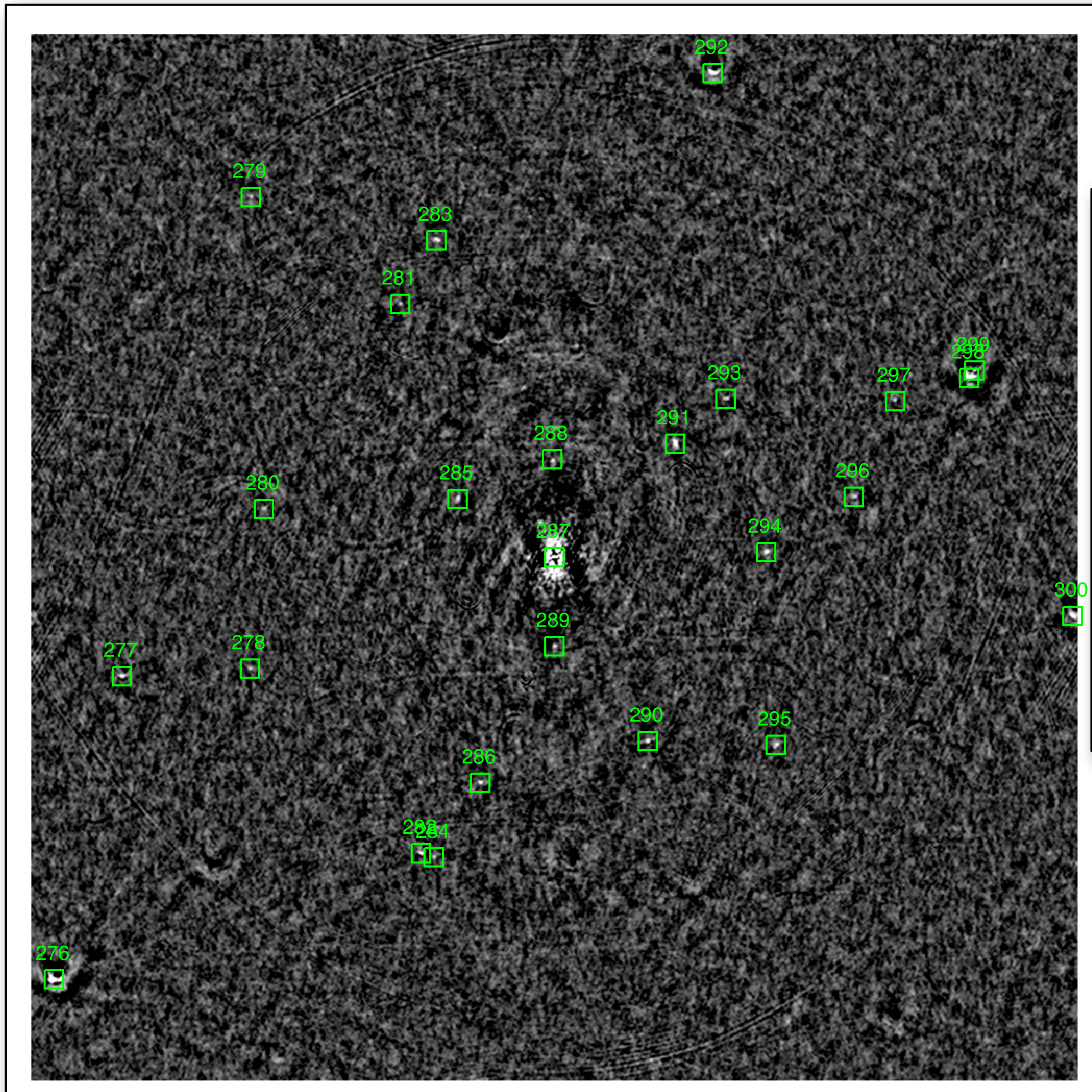
(courtesy S. Yatawatta)

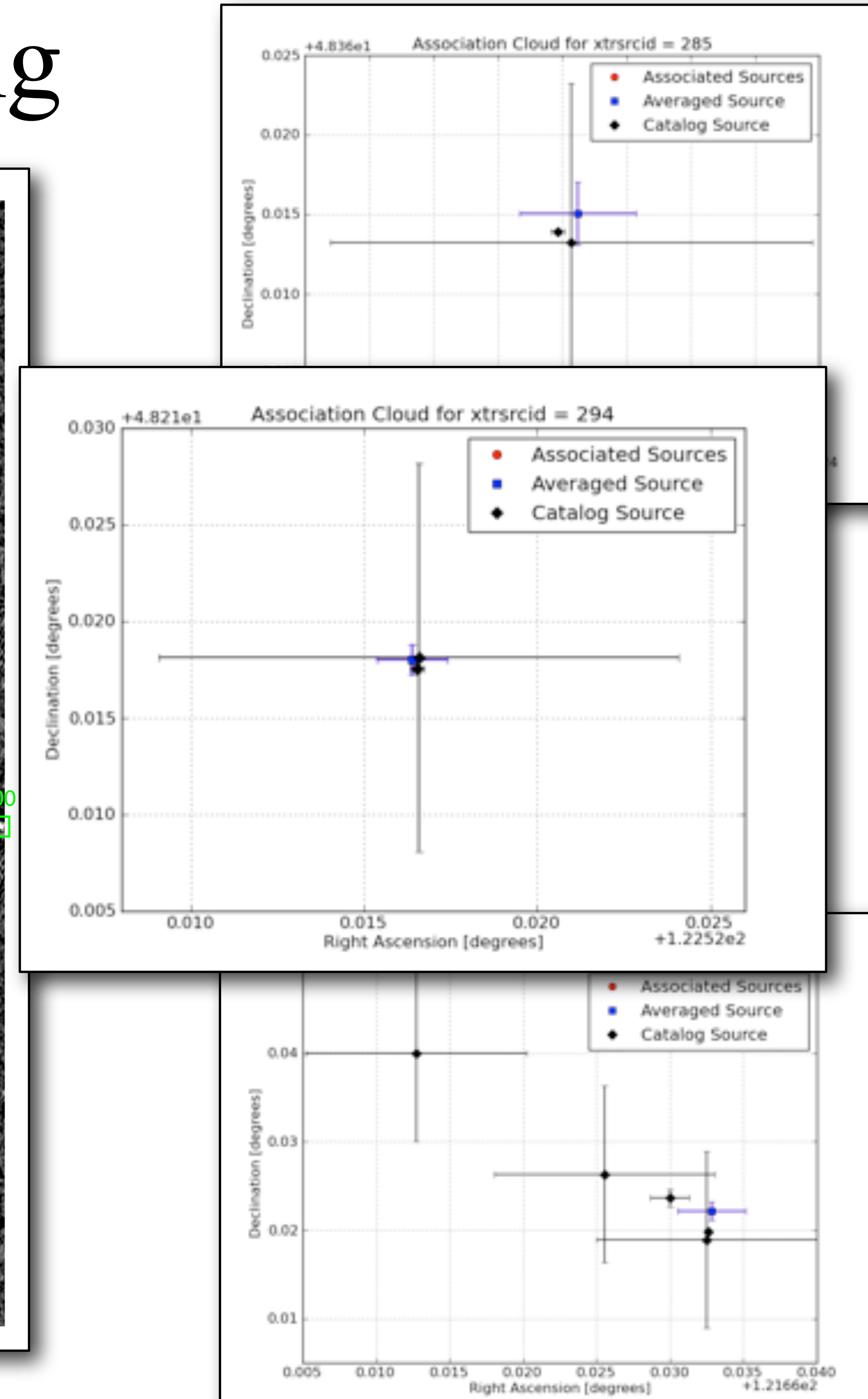
3C 61.1
Wide-field imaging

HBA 115-185 MHz
8(x2)+ 4 stations
8 deg x 8 deg field
4 arcsec pixels
~5.18×10^{7} pixels
10 arcsec PSF

10 Jy peak
1 mJy noise

GSM and Source Finding





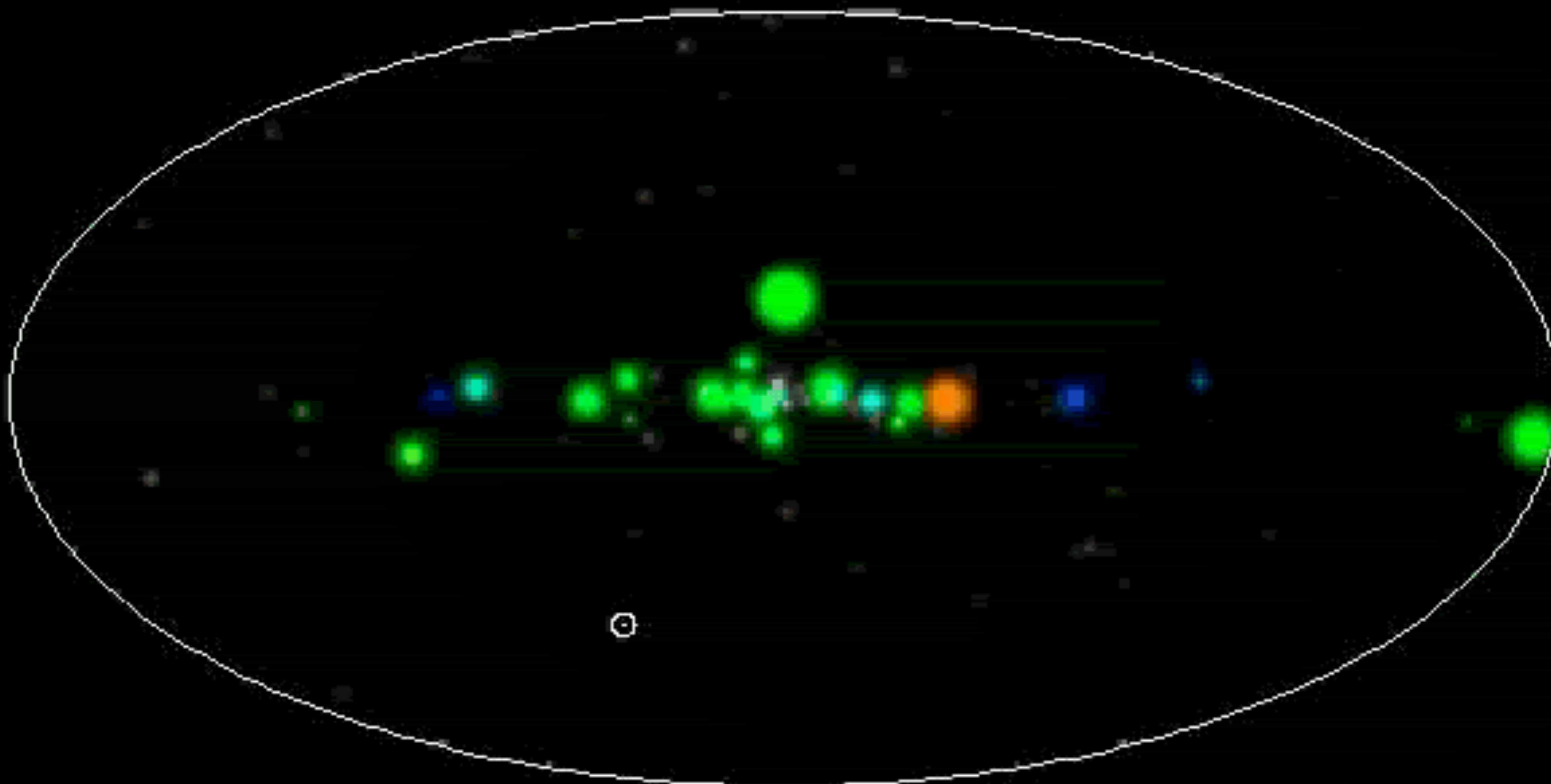
(courtesy B. Scheers, N. Mohan, J. Swinbank)

Catalog Sizes and Data Volumes

ν MHz (1)	Confusion level mJy (2)	Source density arcmin^{-2} (3)	Number of sources in beam (4)	Integration time to reach confusion hour (5)	Total number of sources all sky (6)
15	4.745	0.2	2.7e+05	48	1.3e+07
30	0.969	0.7	2.7e+05	38	5.4e+07
60	0.205	2.9	2.7e+05	585	2.2e+08
75	0.124	4.5	2.7e+05	991	3.4e+08
120	0.043	11.6	2.7e+05	23	8.6e+08
150	0.026	18.1	2.7e+05	55	1.3e+09
200	0.014	32.2	2.7e+05	191	2.4e+09
240	0.009	46.3	2.7e+05	668	3.4e+09

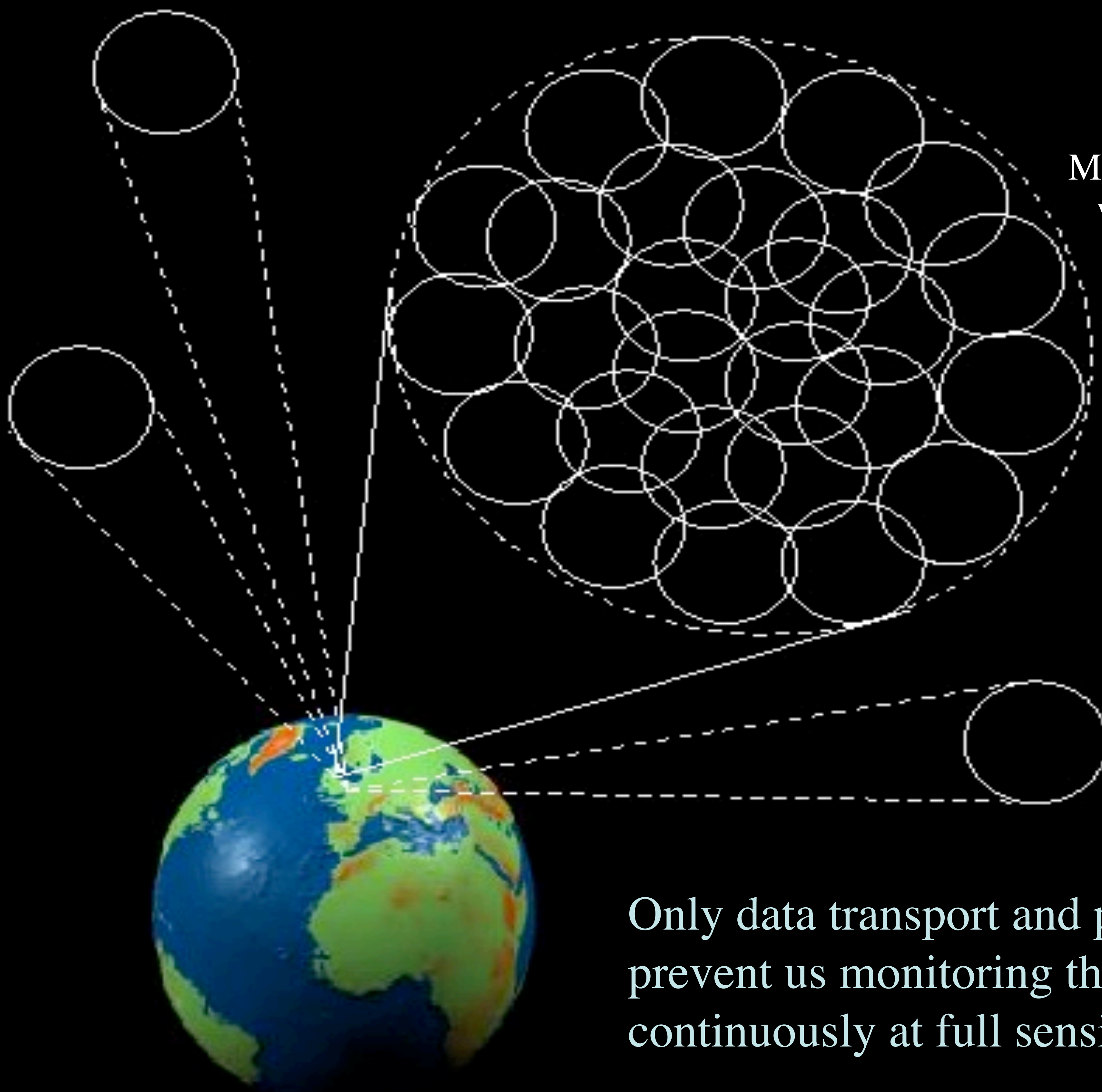
ν MHz (1)	Number of pixels (2)	Amount of data continuum maps Gbyte (3)	Amount of data spectral data cubes at full polarization Tbyte (4)
15	4.2e+08	1	27
30	1.7e+09	6	108
60	6.8e+09	27	432
75	1.1e+10	42	675
120	2.7e+10	108	1729
150	4.2e+10	168	2702
200	7.5e+10	300	4803
240	1.1e+11	432	6917

The RXTE All-Sky Monitor Movie



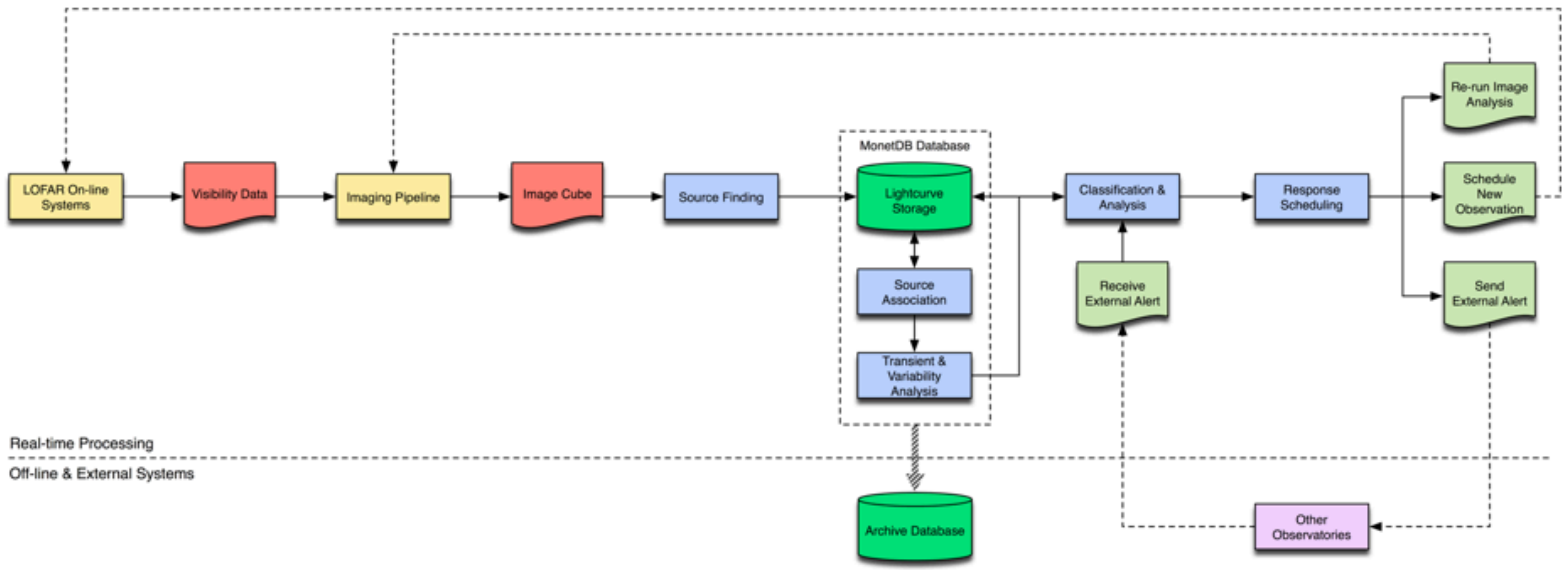
02 / 23 / 1996

Radio Sky Monitor: Multiple station beams tile out a significant fraction of the sky and detect transient sources on timescales down to 1 second



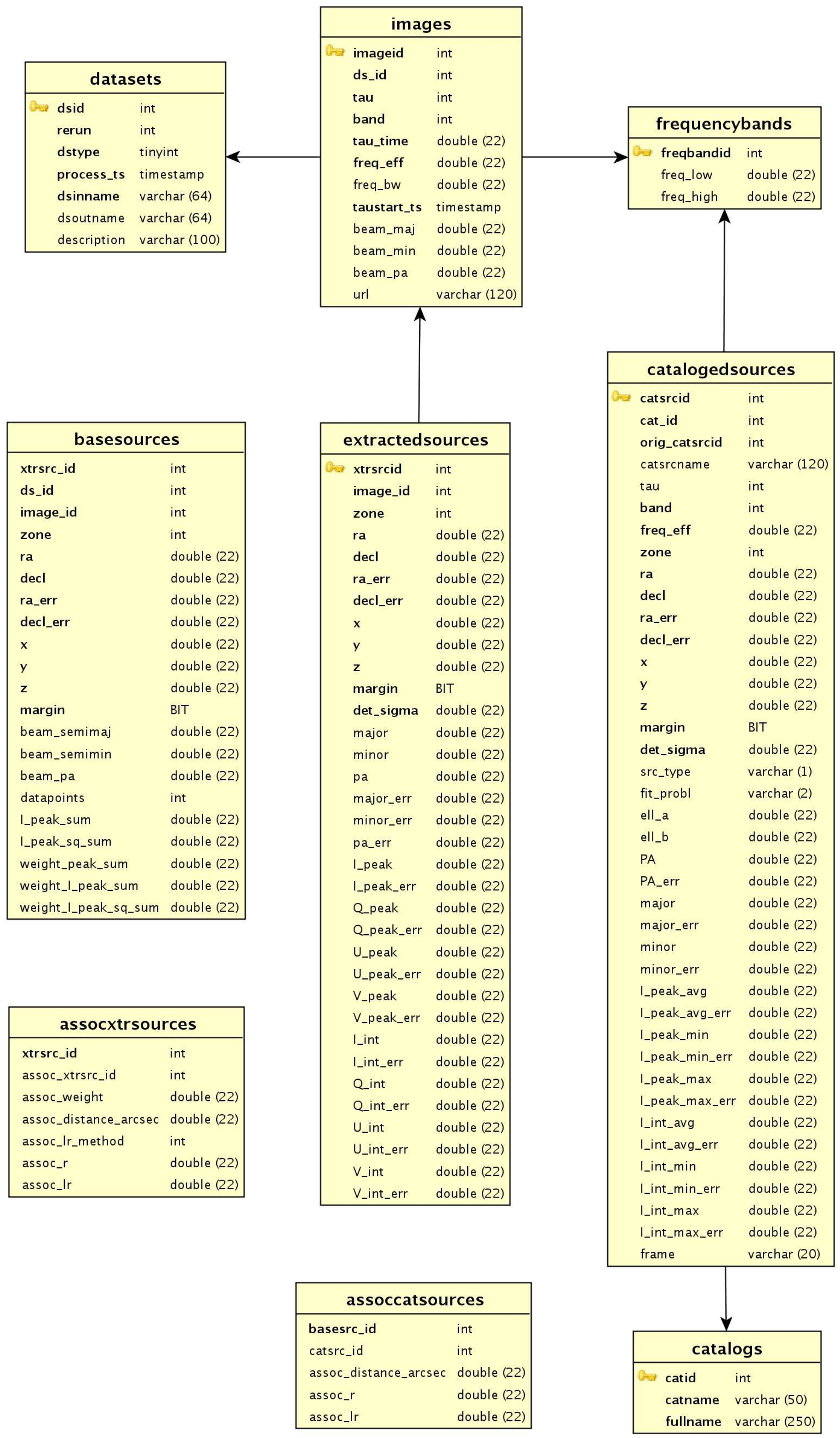
Only data transport and processing power prevent us monitoring the entire visible sky continuously at full sensitivity and resolution!

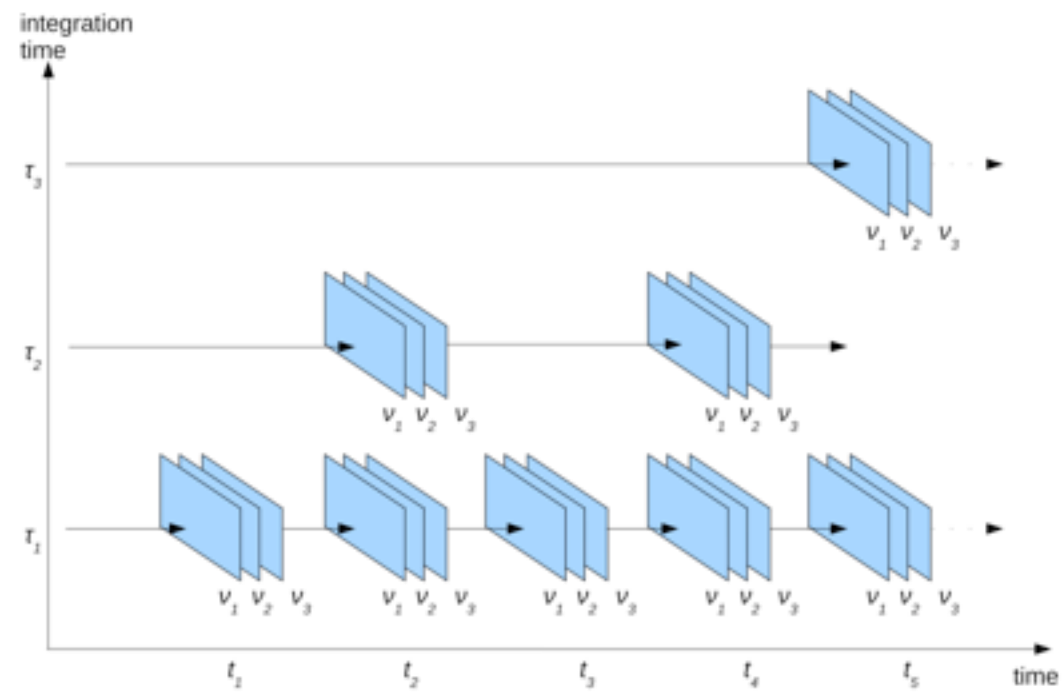
Transient Detection Pipeline

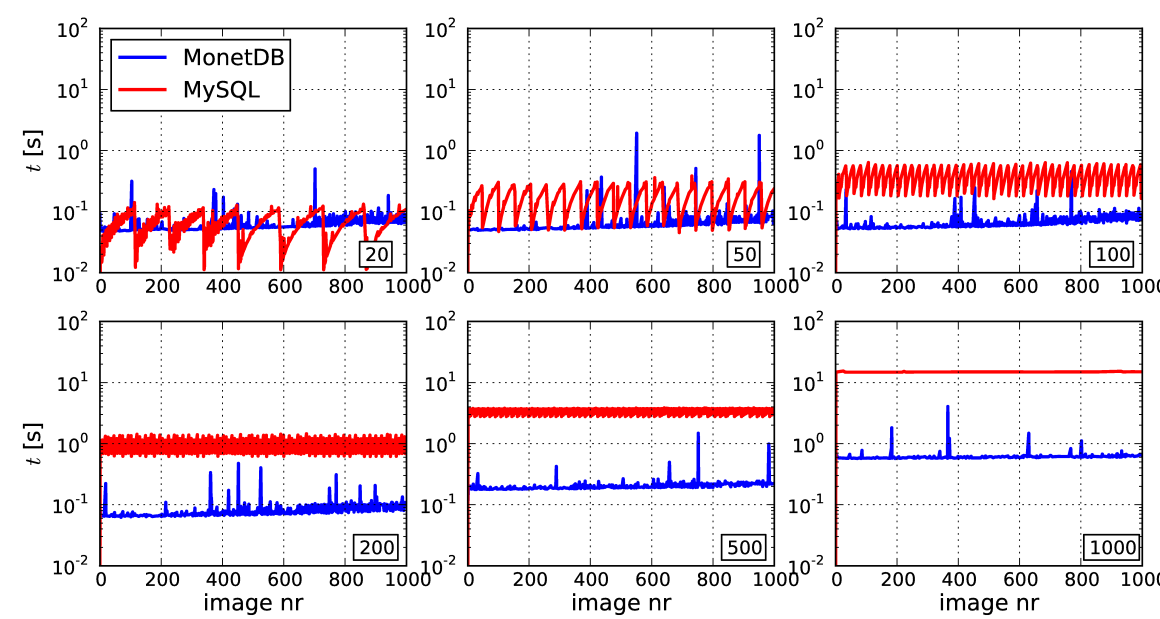


Expansion on standard imaging mode
Requirement for near real-time performance
Detection, classification, and response
Generate and receive event triggers (internal and VOEvents)

Transient Database Design

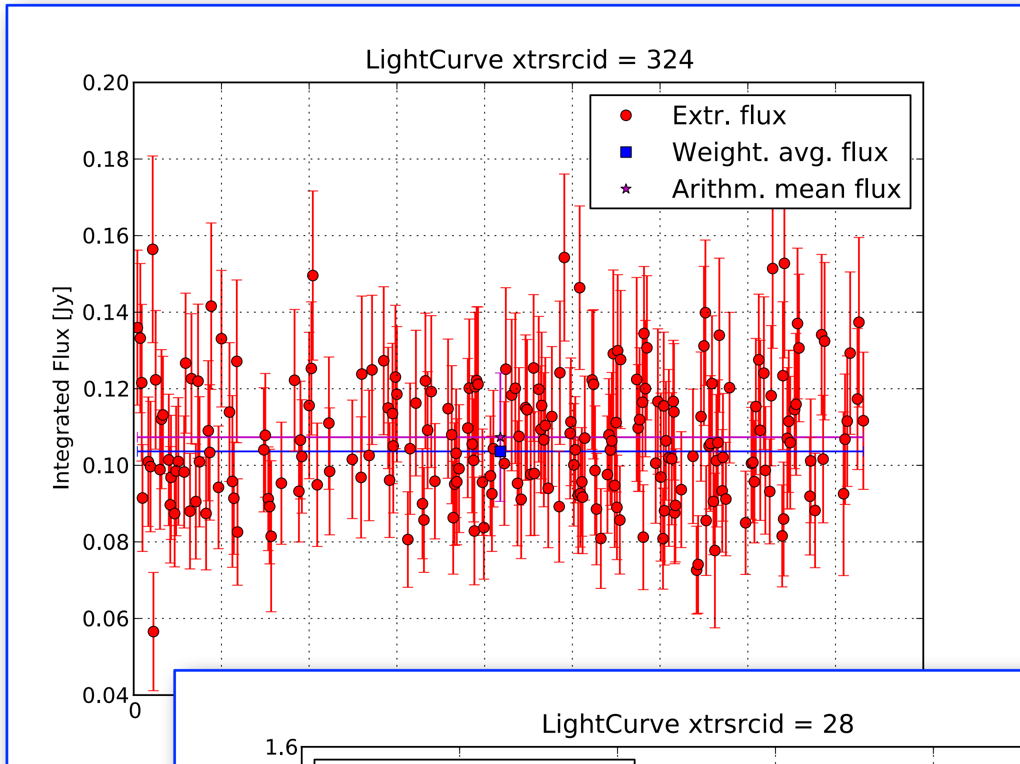


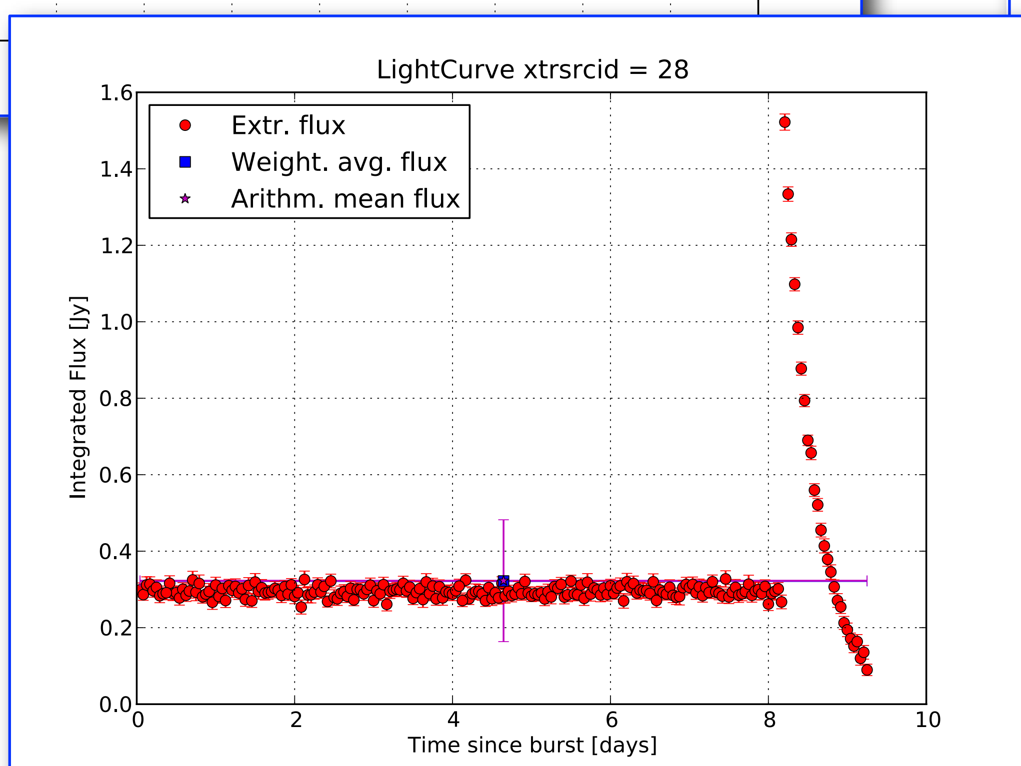


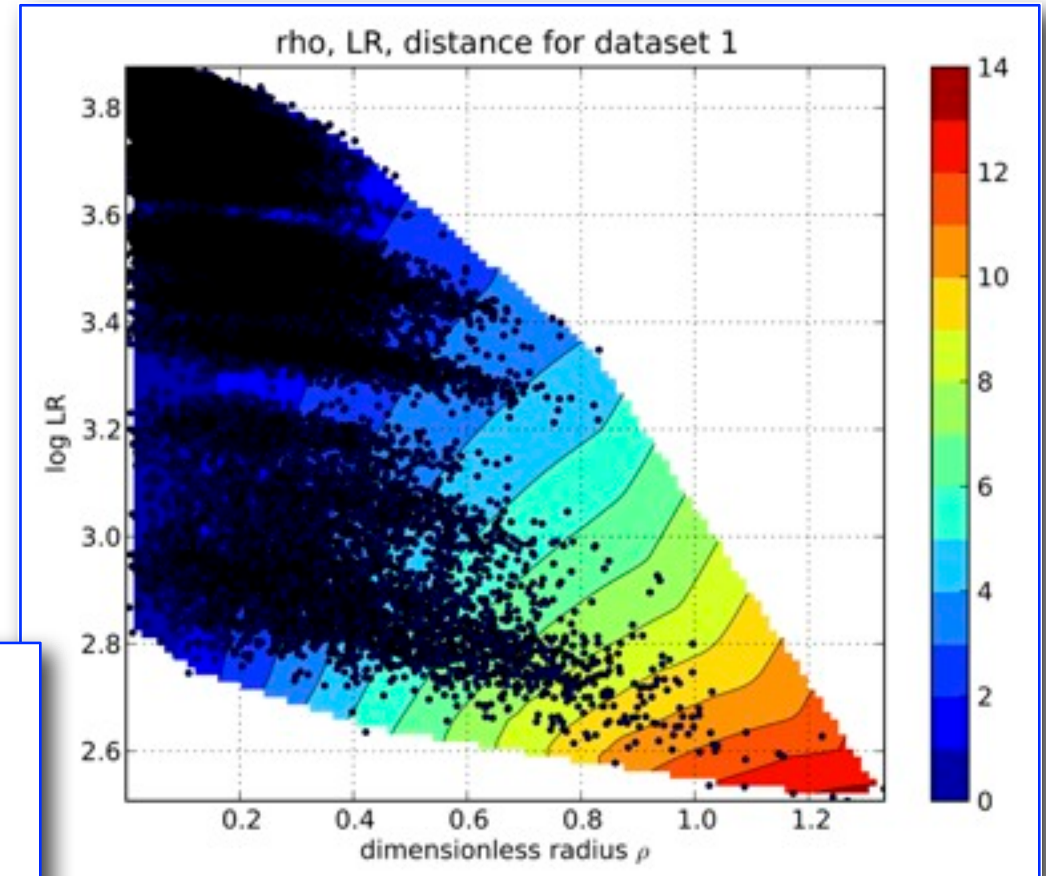


(courtesy B. Scheers)

Transient Detection Pipeline





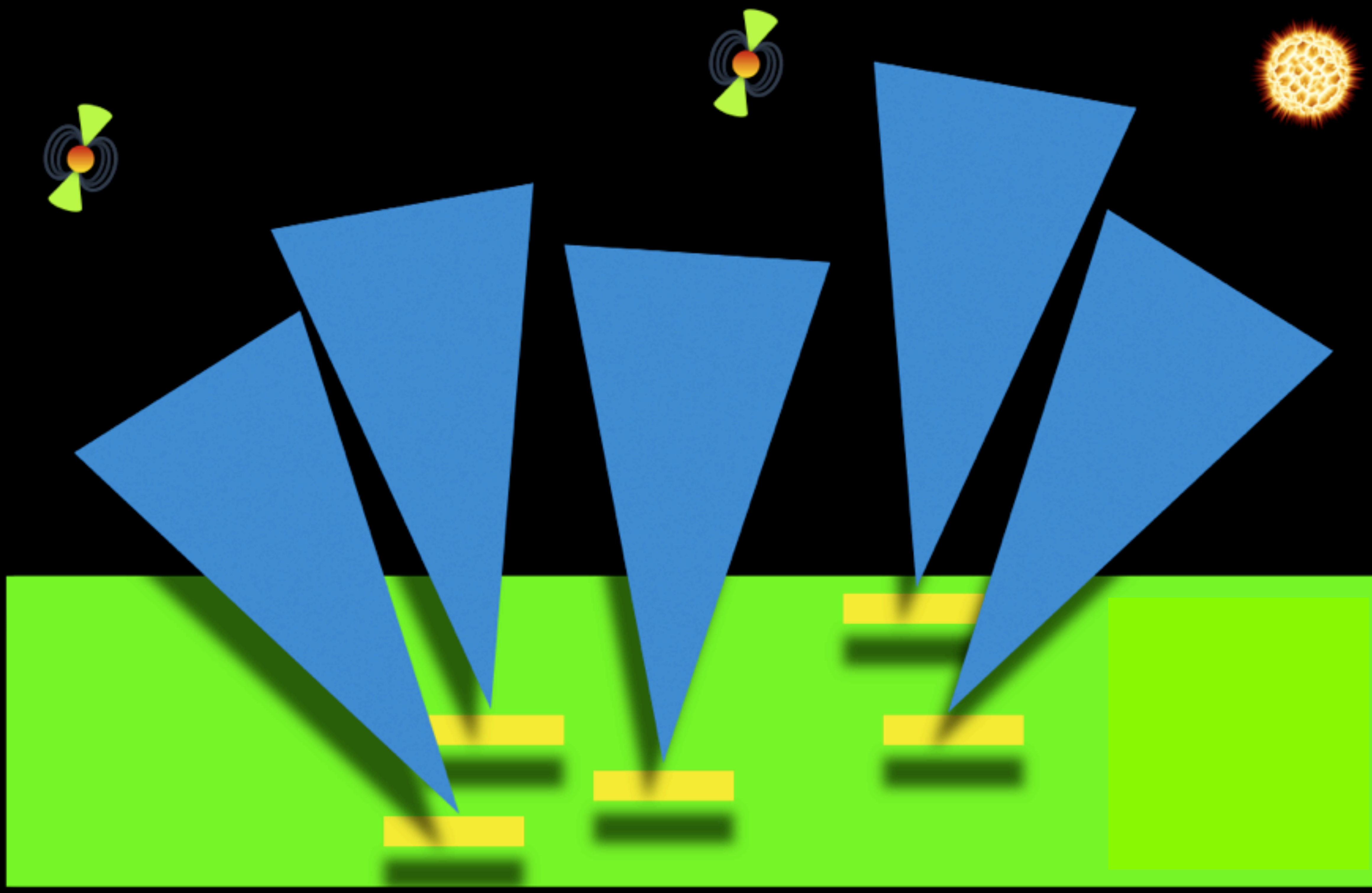


(courtesy B. Scheers)

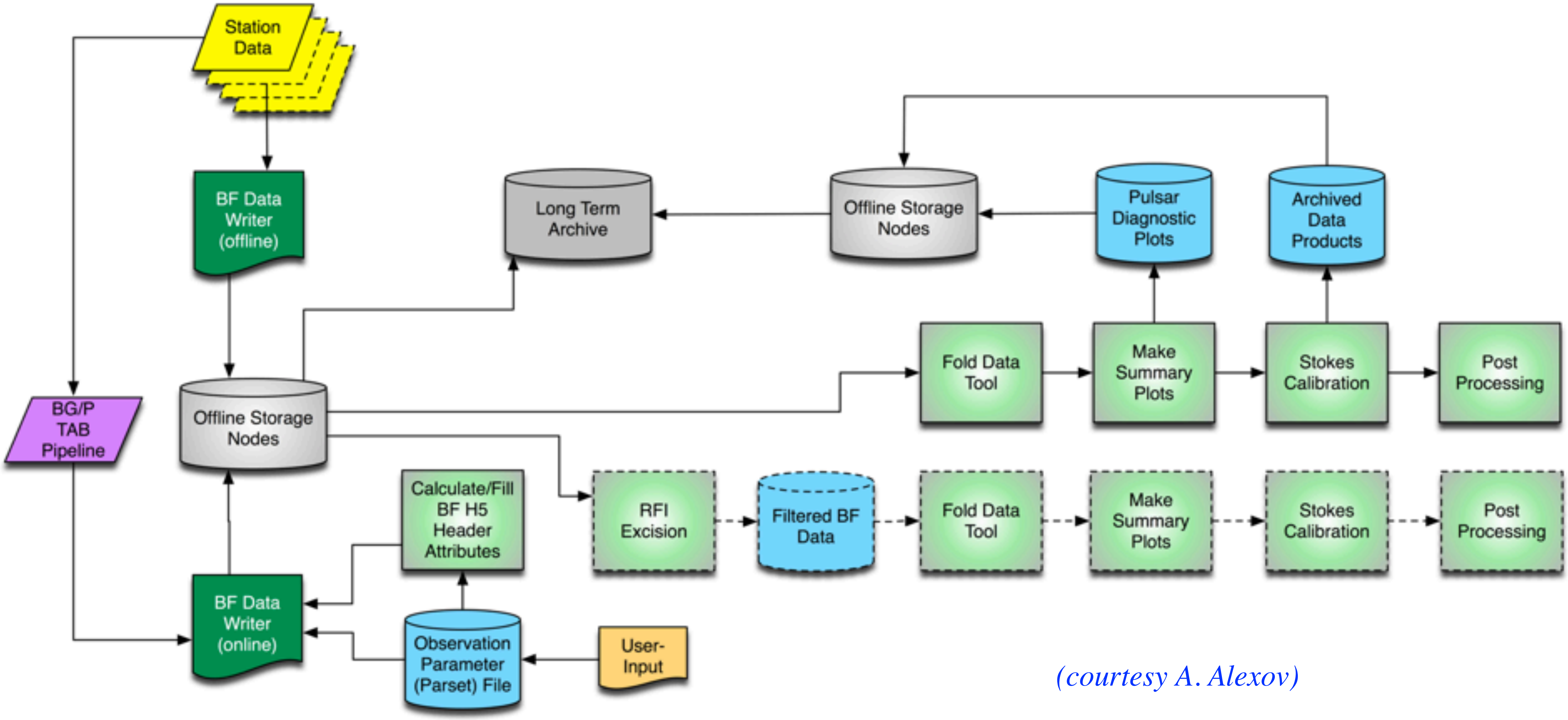
Sample extracted lightcurves

Push computations to the DB!

Pulsar Surveys with LOFAR

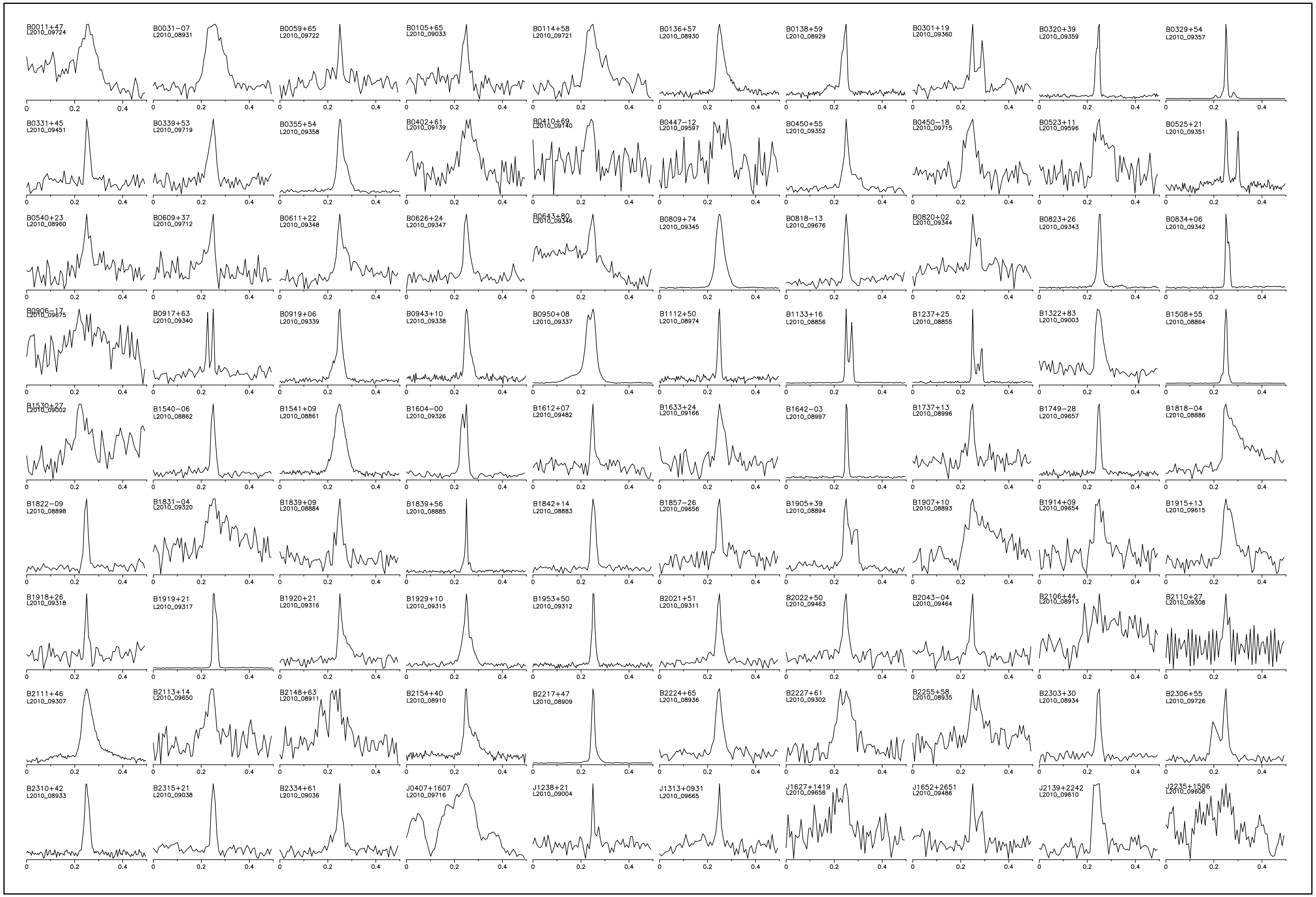


Known Pulsar Pipeline



(courtesy A. Alexov)

100+ Pulsars Detected with LOFAR

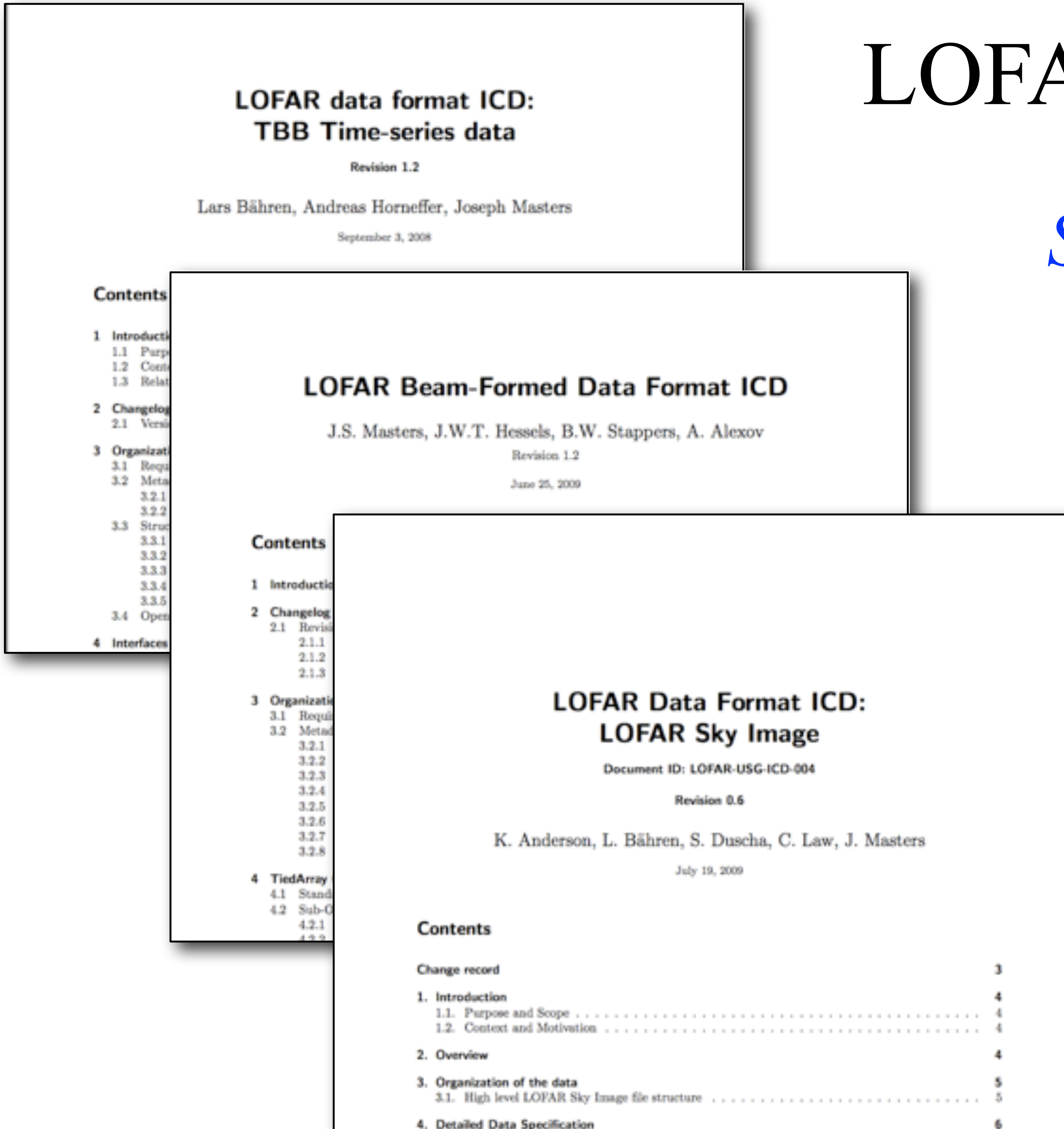


(courtesy: J. Hessels & Pulsars WG)

Beam-formed Data Rates

Mode	Description	Data Rate	FoV (sq. deg.)	Res. (deg.)	Sens. (norm.)
Incoherent (par. imaging)	Stations added without proper phase correction.	2-250 GB/hr	12.5	2	6.0
Tied-array	Stations added properly in phase.	Up to 23TB/hr	0.2	0.03	36.0
Single Station	For projects with high time, but lower sensitivity requirements.	2-250 GB/hr	12.5	2	1.0
Superstation	Interesting balance of sensitivity and FoV.	Up to 23TB/hr	9.0	0.2	12.0
Fly's Eye	Maximize total FoV for bright transient survey.	Up to 8TB/hr	450	2	1.0

LOFAR Data Formats


LOFAR data format ICD:
TBB Time-series data
Revision 1.2
Lars Bähren, Andreas Horneffer, Joseph Masters
September 3, 2008
Contents

LOFAR Beam-Formed Data Format ICD
J.S. Masters, J.W.T. Hessels, B.W. Stappers, A. Alexov
Revision 1.2
June 25, 2009
Contents

LOFAR Data Format ICD:
LOFAR Sky Image
Document ID: LOFAR-USG-ICD-004
Revision 0.6
K. Anderson, L. Bähren, S. Duscha, C. Law, J. Masters
July 19, 2009

Contents

Sky Cubes

BF Data Products

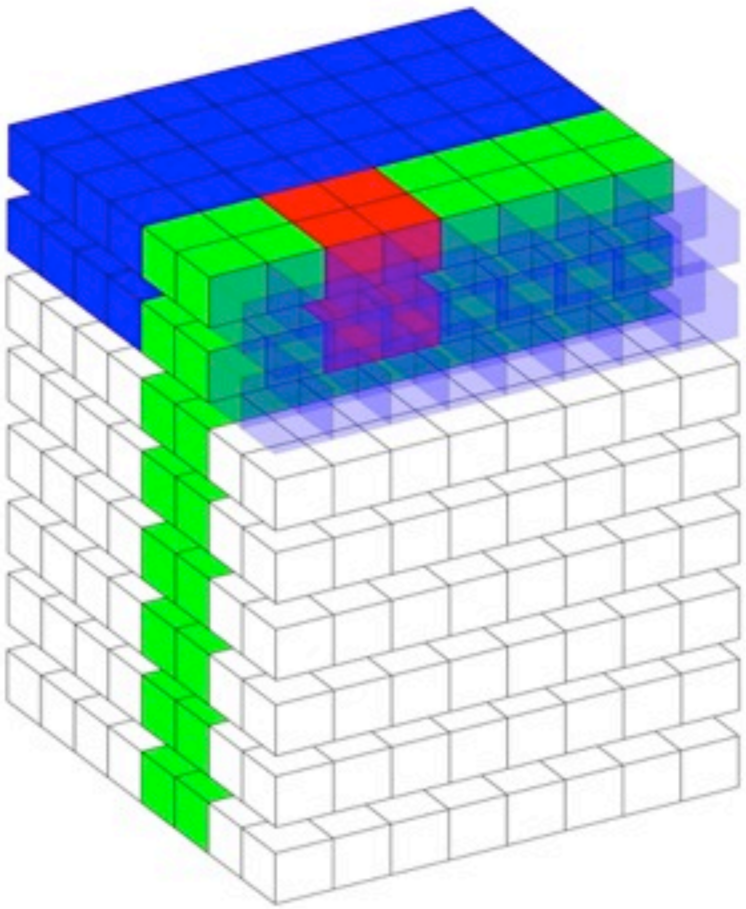
TBB Time Series

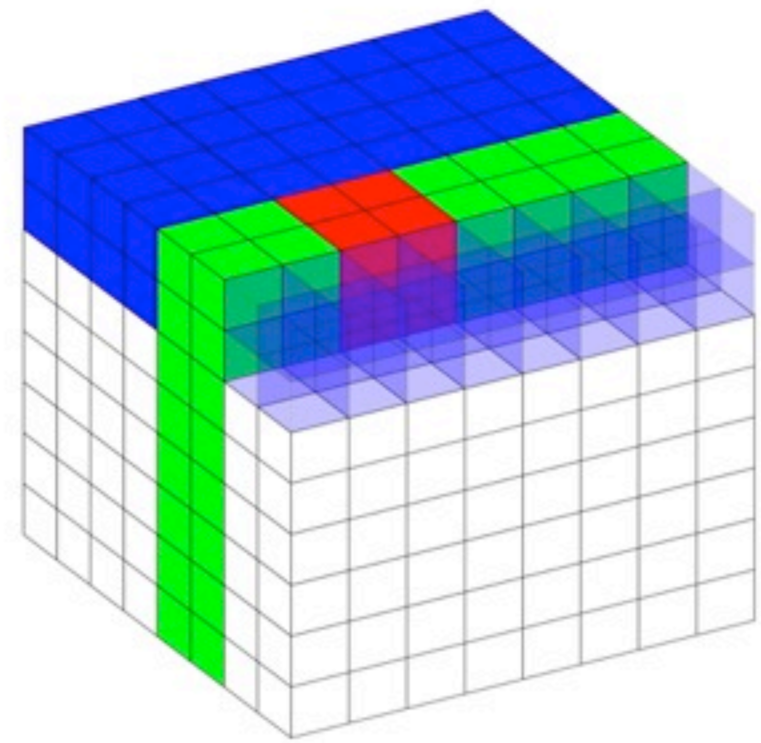
Near-field Cubes

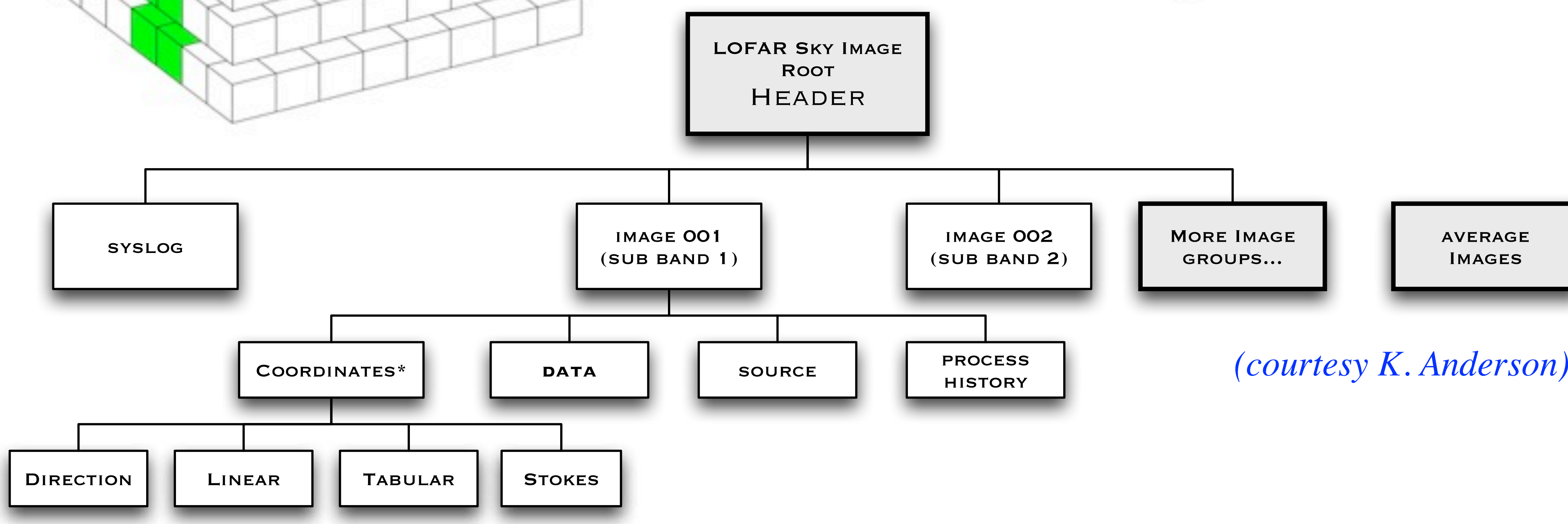
Dynamic Spectra

RM Cubes

LOFAR HDF5 Sky Cubes

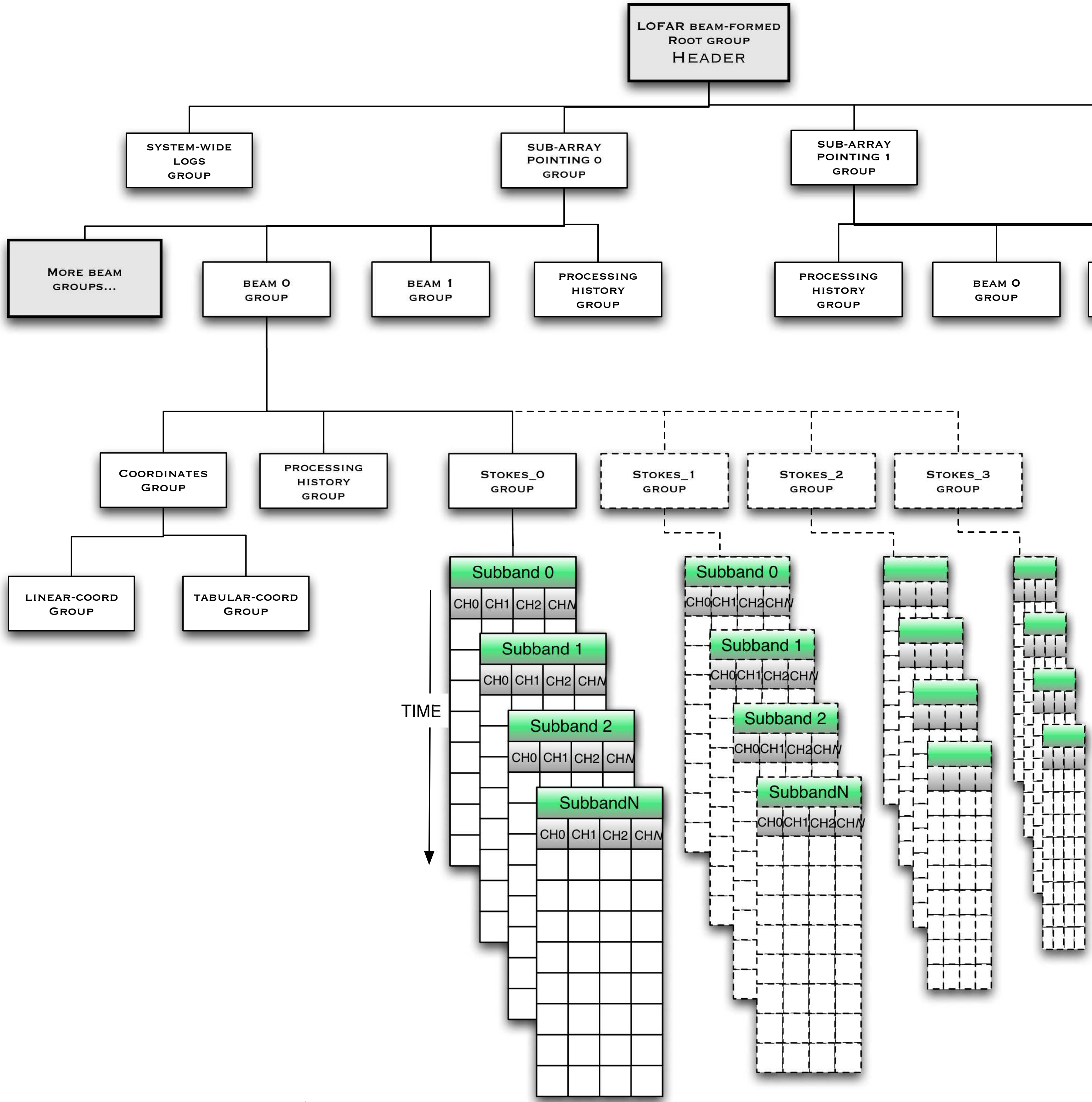






(courtesy K. Anderson)

LOFAR HDF5 Beam-formed Data

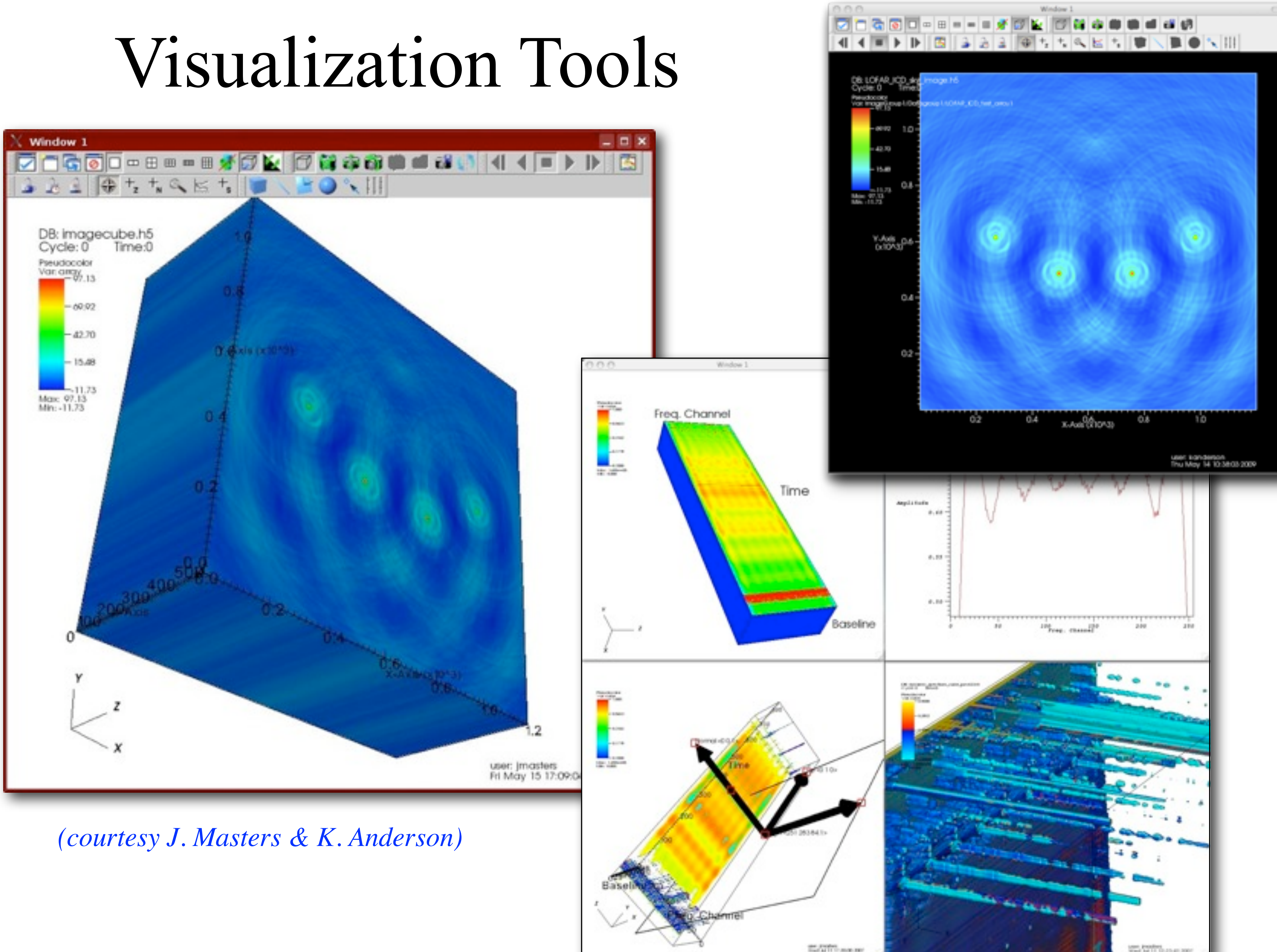


Data Packaging Options:

1) One Table per Subband (1d) [Shown here]
2) One Table for all Subbands (Nd)
3) 1 Array per Subband (1d)
4) 1 Array for all Subbands (Nd)

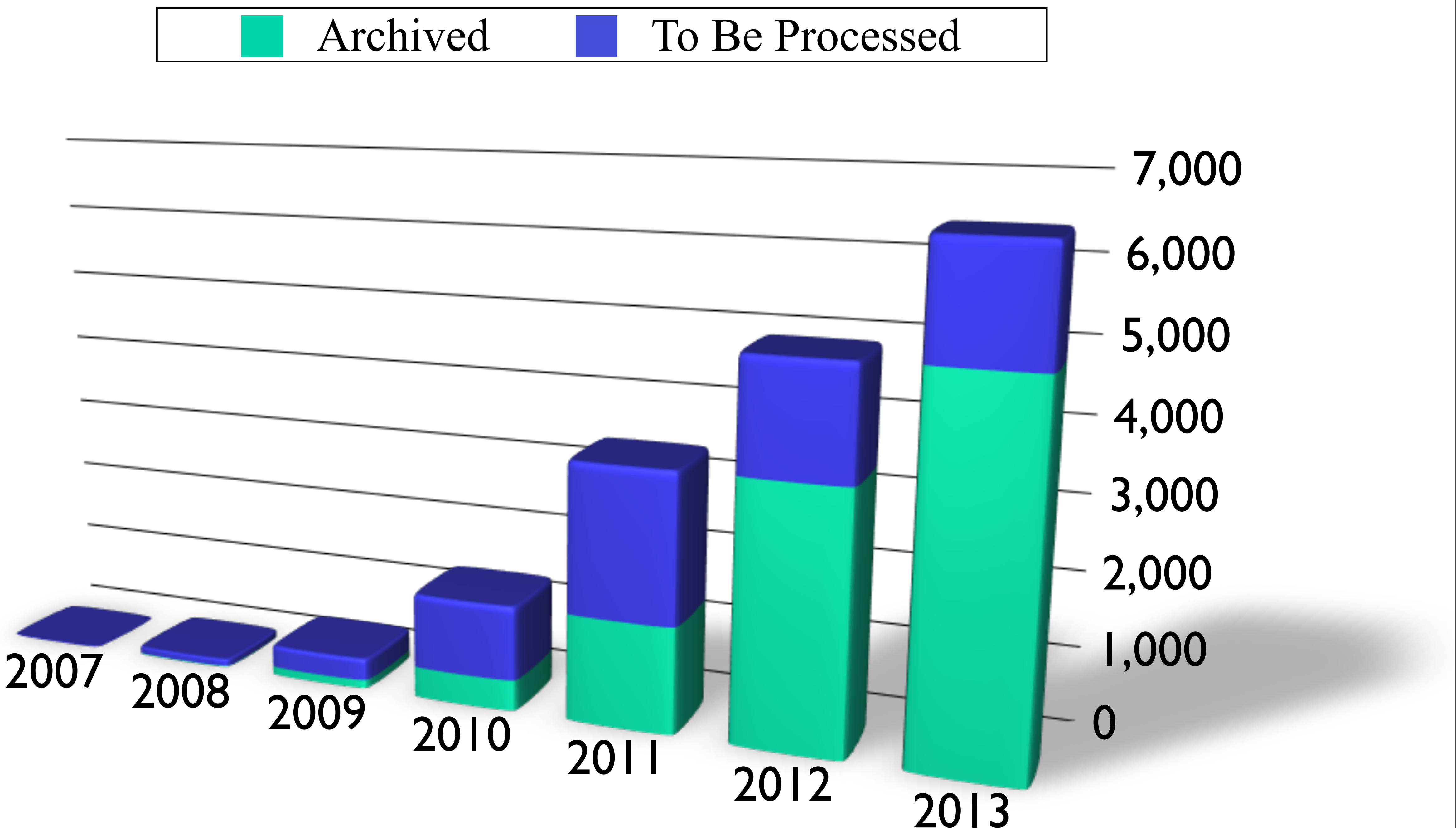
(courtesy A. Alexov)

Visualization Tools



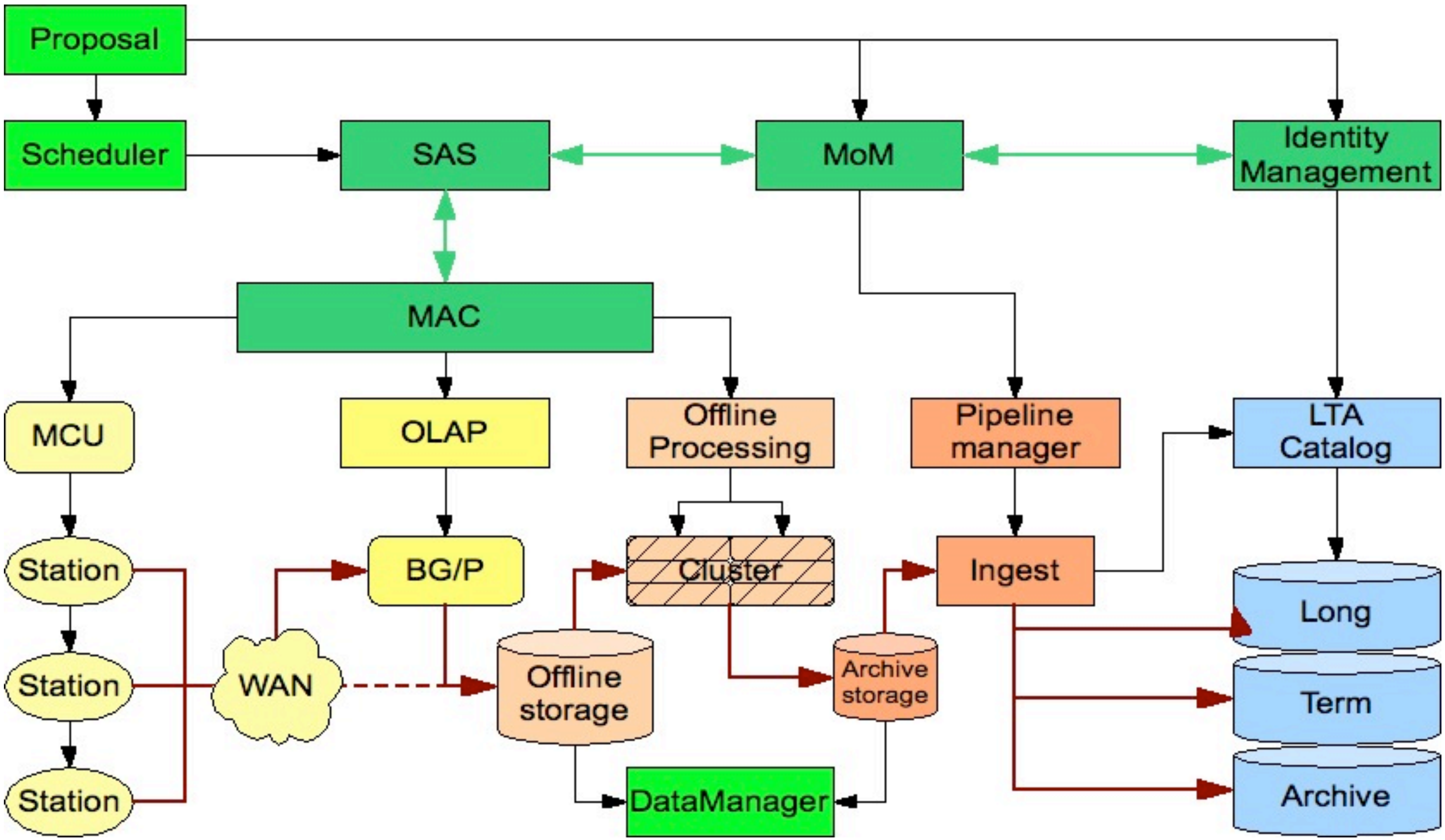
(courtesy J. Masters & K. Anderson)

LOFAR Archive Estimates

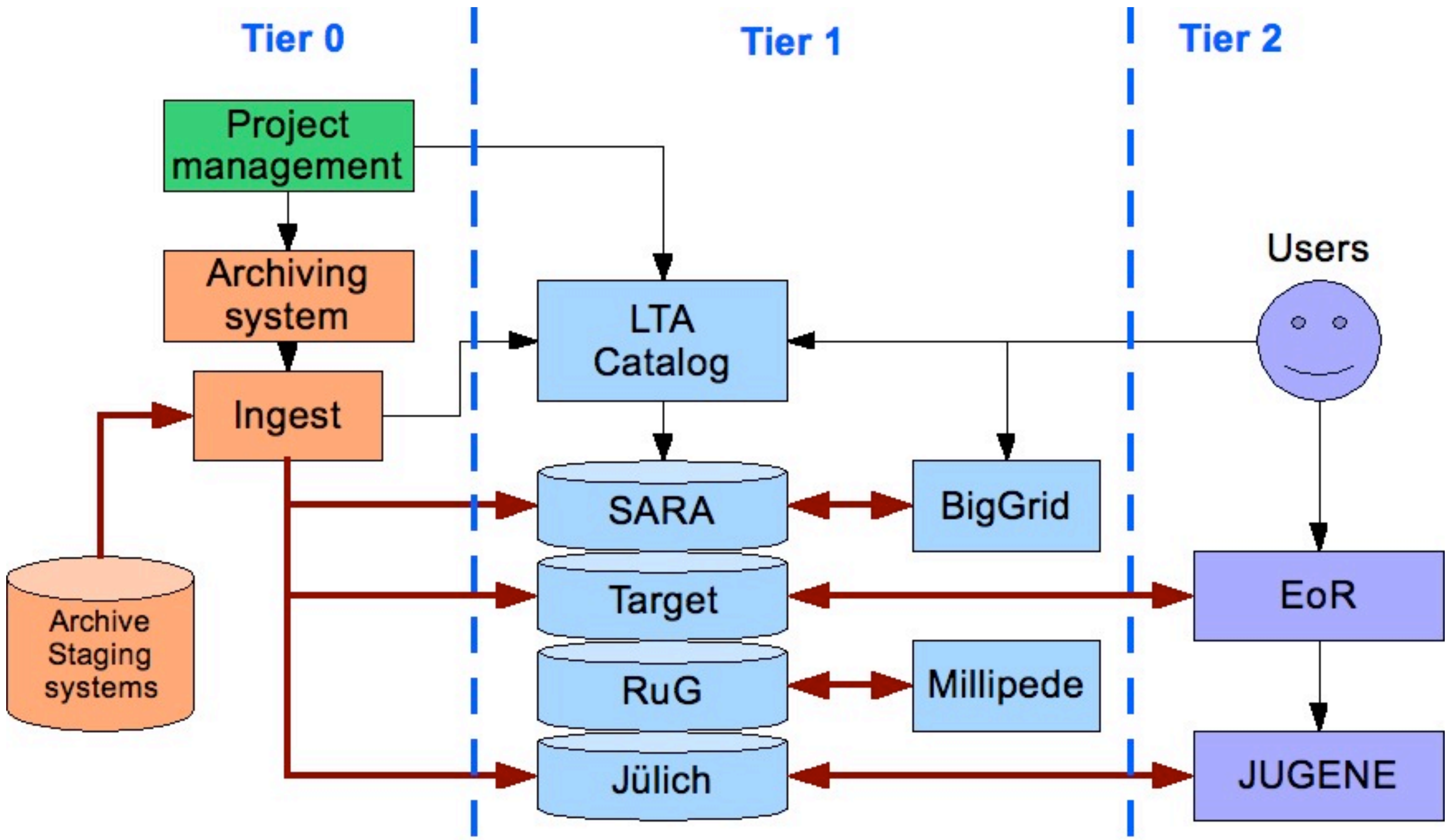


Estimated growth rate ~ 2.5 Pb/yr

LOFAR Central Processing



LOFAR Archive Topology



Current Archive Capacity

	Online (disk)	Nearline (tape)	Used
SARA	300 TB	1 PB	250 TB
Jülich	90 TB	1 PB	10TB
Target	1+ PB	3+ PB	-

Target systems will become operational in about May 2011

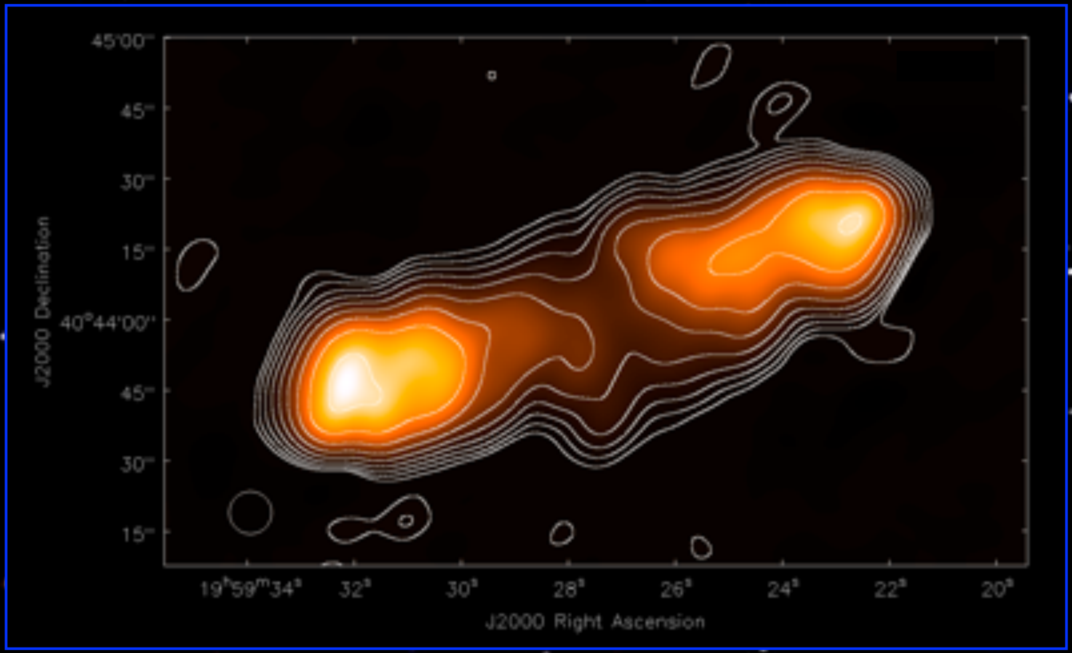
	Cores
SARA (BigGrid)	2200
NIKHEF (BigGrid)	2200
RuG (BigGrid)	770
Jülich (JUROPA)	17500

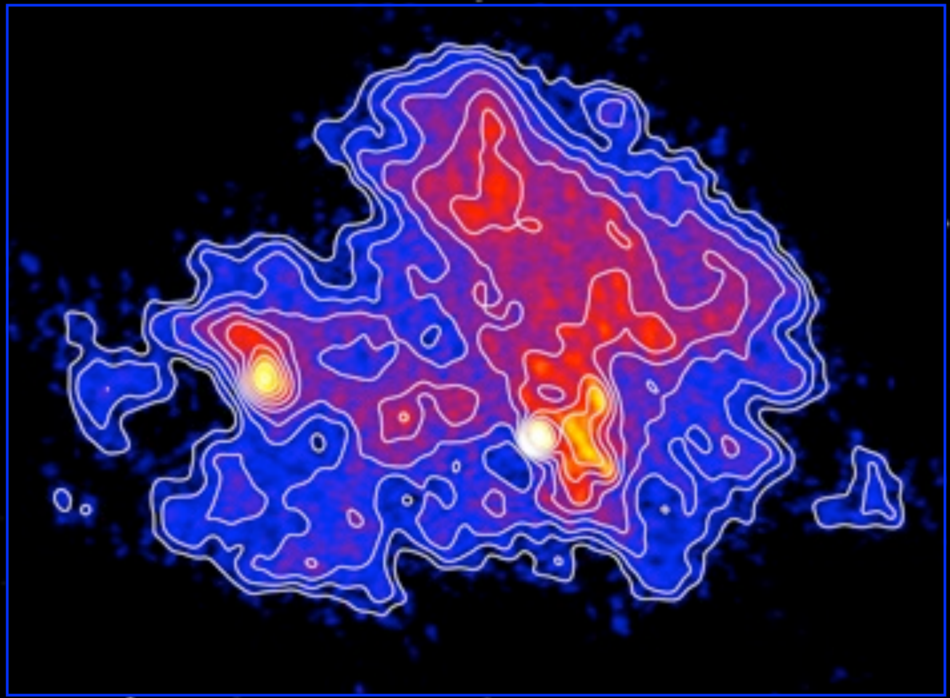
These are shared compute clusters

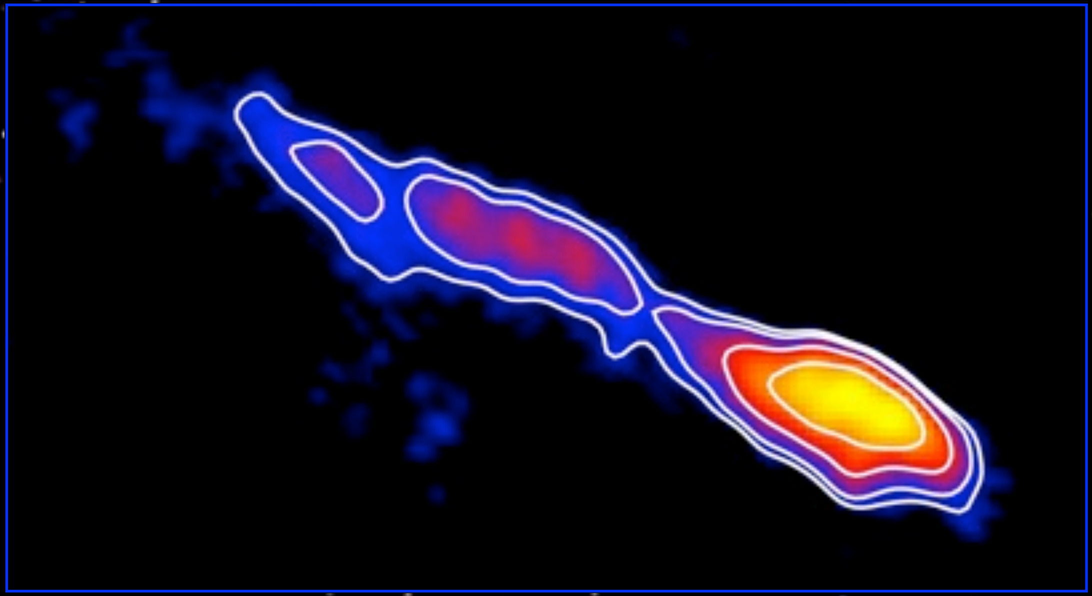
Summary

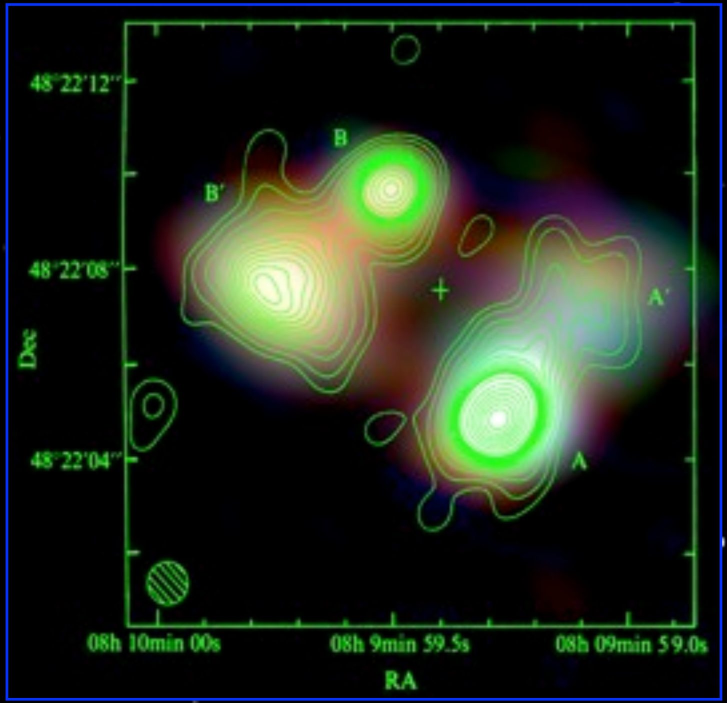
- **Hardware roll-out complete early 2012**
- **Science pipelines under continuing development**
- **Heavy commissioning throughout 2011-2012**
- **Data volume and management is already an issue**
- **Data management effects processing strategies**
- **Requires trade-offs between quality and efficiency**
- **Real-time science drivers require high performance**
- **Pipelines produce a zoo of large and complex datasets**
- **Data management will drive archive content**
- **Archives must become processing centers**

The End


45'00"
45"
30"
15"
40°44'00"
45"
30"
15"
J2000 Declination
19h59m34s
32s
30s
28s
26s
24s
22s
20s
J2000 Right Ascension






48°22'12"
48°22'08"
48°22'04"
Dec
B
B'
A
A'
08h 10min 00s
08h 9min 59.5s
08h 09min 59.0s
RA