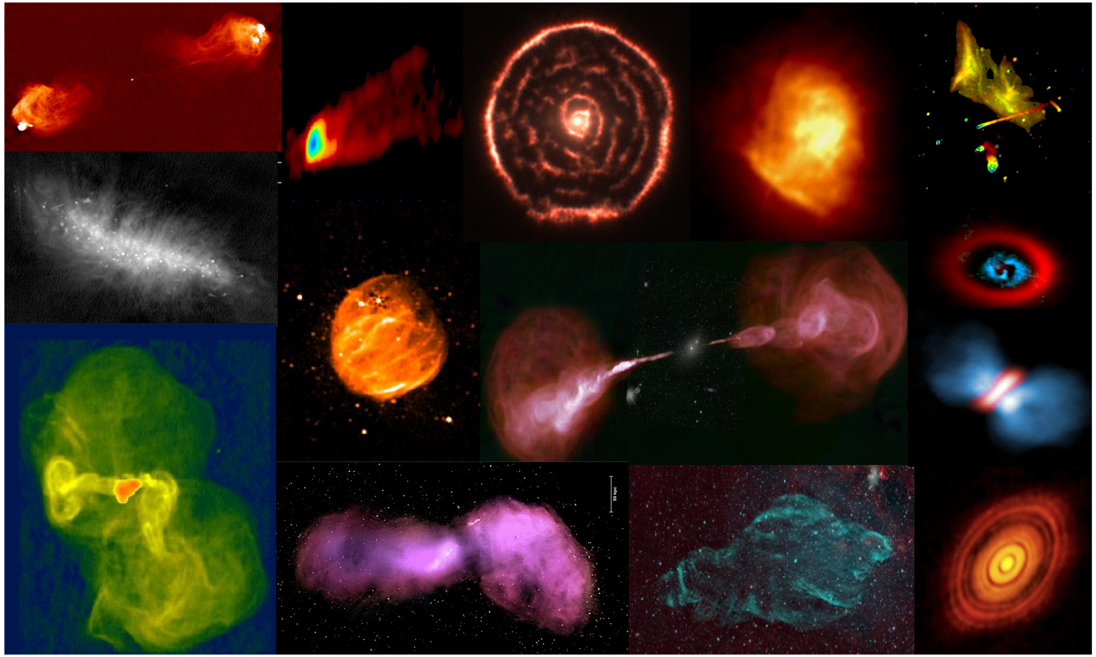
Overview : Radio Interferometric Data Analysis and Compute Needs



Urvashi Rau
National Radio Astronomy Observatory, Socorro, NM, USA

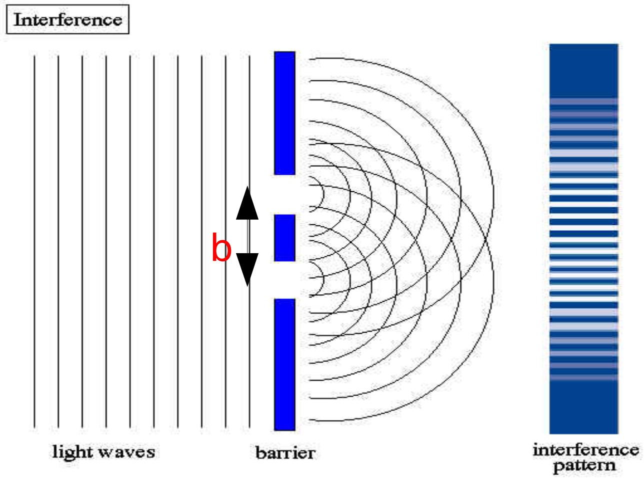
25th March 2022
HPC Workshop on Radio Astronomy Data Analysis in the SKA Era
40th Meeting of the Astronomical Society of India

Outline

- Introduction to Radio Interferometry
- Data Management
  - Data Acquisition
  - Flagging, Calibration, Imaging
  - Pipelines and Automation
- Areas of HPC application and innovation

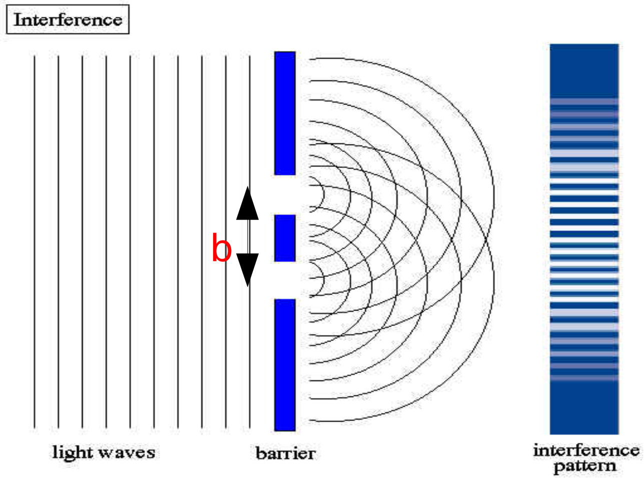
An indirect imaging device

Young’s double slit experiment



An indirect imaging device

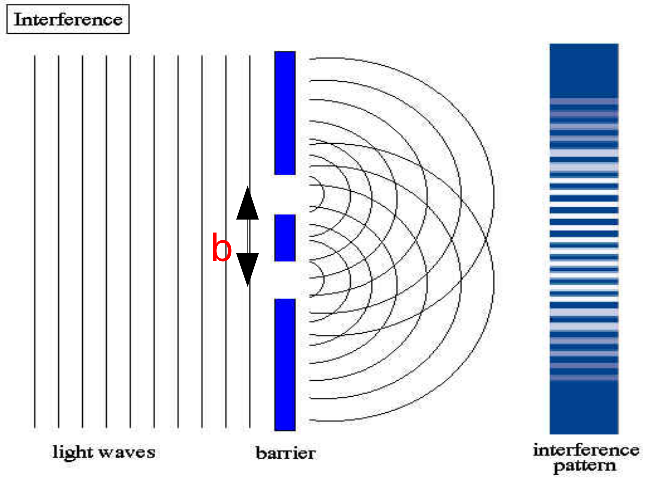
Young's double slit experiment



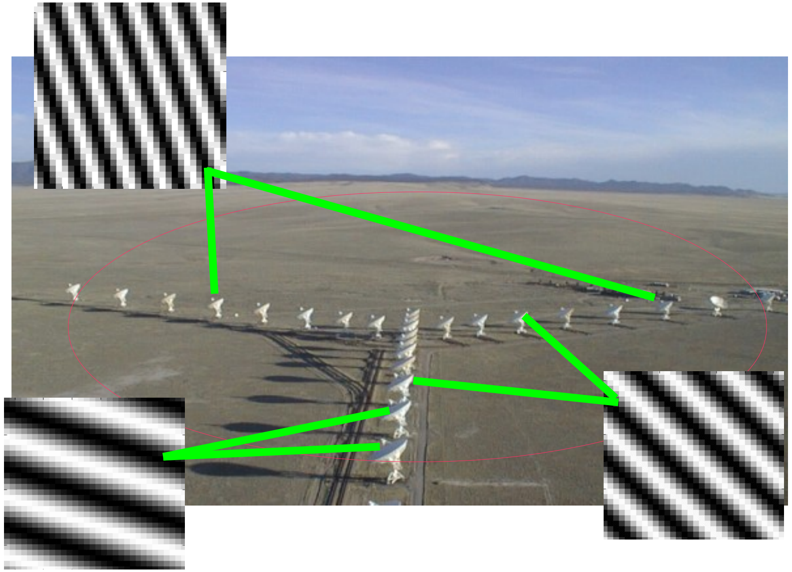
Instrument : An array of detectors

An indirect imaging device

Young’s double slit experiment



Each antenna-pair measures the parameters of one 'fringe'.

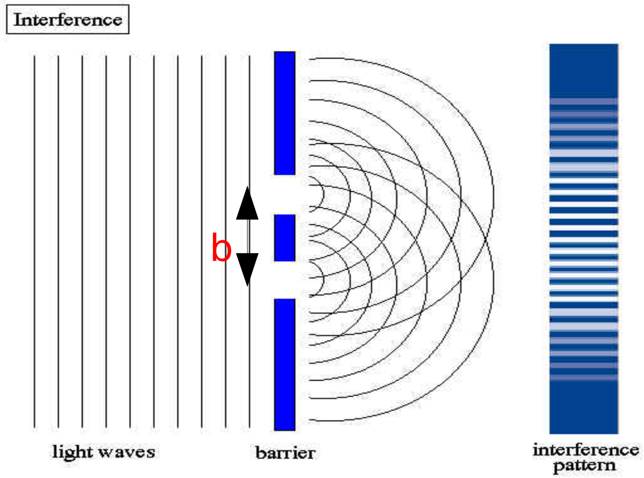


Measured Fringe Parameters :

Amplitude, Phase
Orientation, Wavelength

An indirect imaging device

Young's double slit experiment



2D Fourier transform :

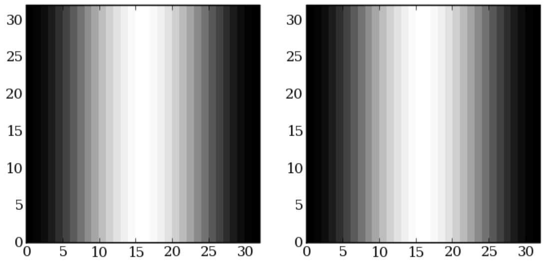
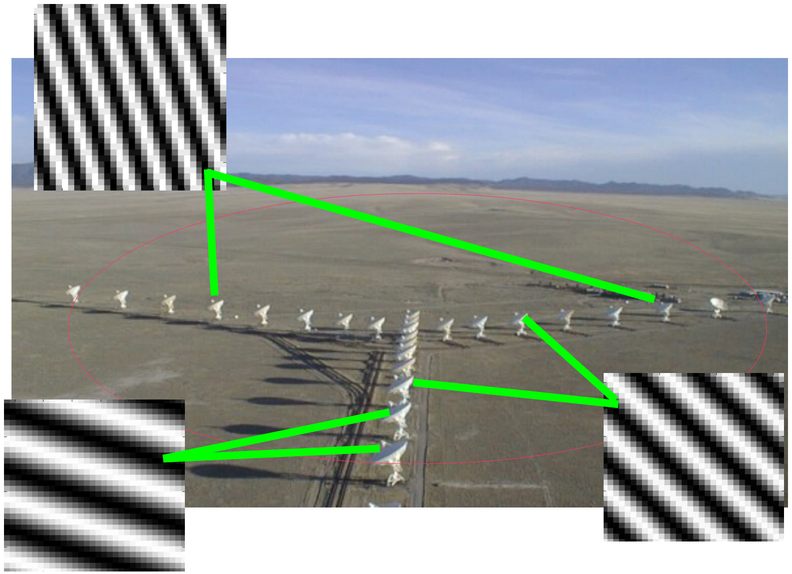


Image = sum of cosine 'fringes'.

Each antenna-pair measures the parameters of one 'fringe'.

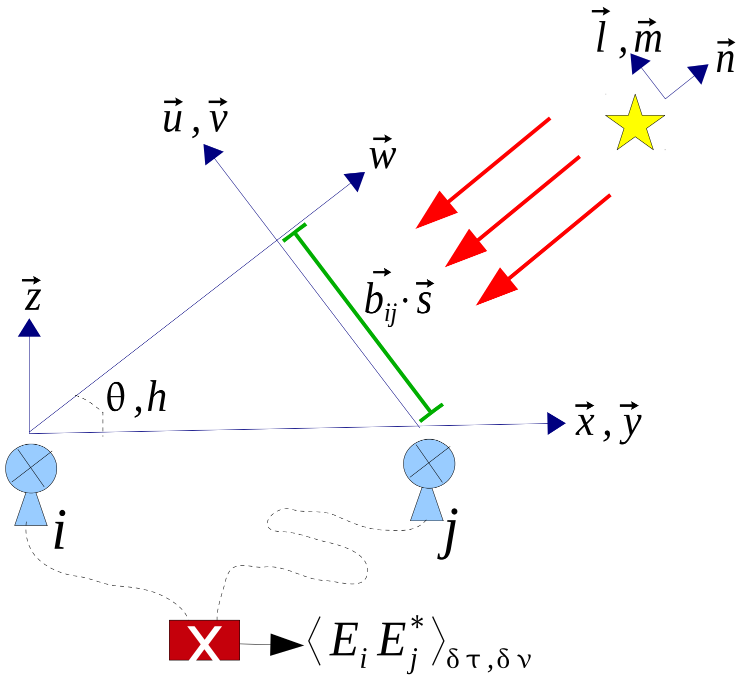


Measured Fringe Parameters :

Amplitude, Phase
Orientation, Wavelength

Measuring the visibility function

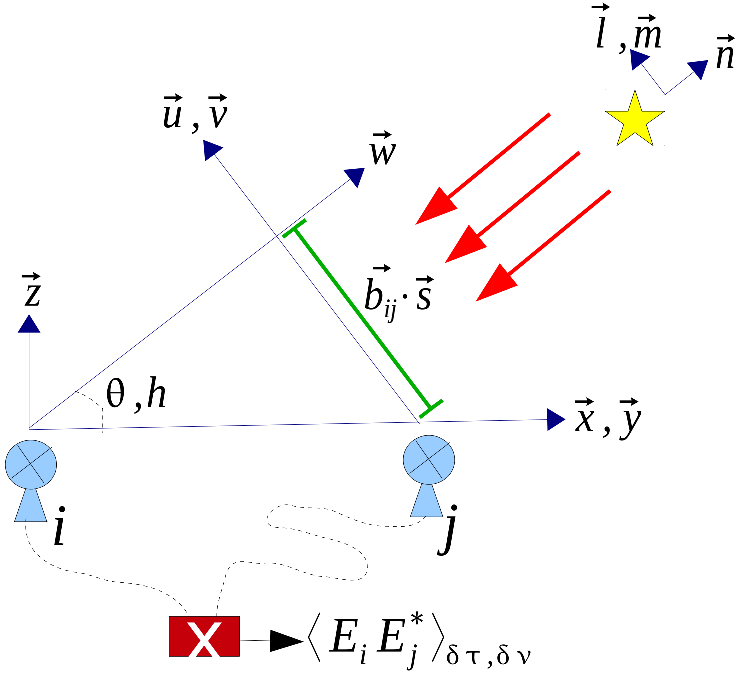
Measure the spatial correlation of the E-field
incident at each pair of antennas



N antennas
N(N-1)/2 antenna-pairs (baselines)

Measuring the visibility function

Measure the spatial correlation of the E-field incident at each pair of antennas



N antennas
N(N-1)/2 antenna-pairs (baselines)

Parameters of a Fringe :

Amplitude, Phase :

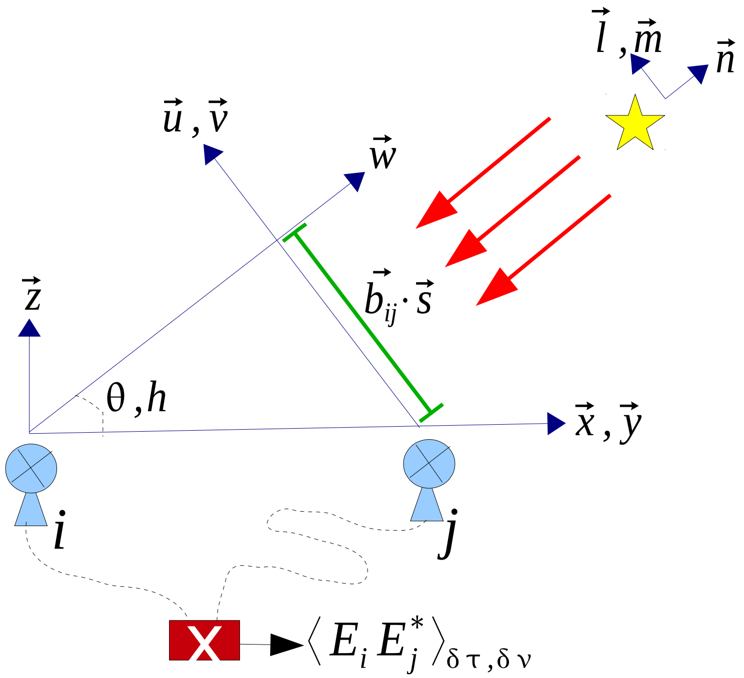
$\langle E_i E_j^* \rangle$ is a complex number.

Orientation, Wavelength :

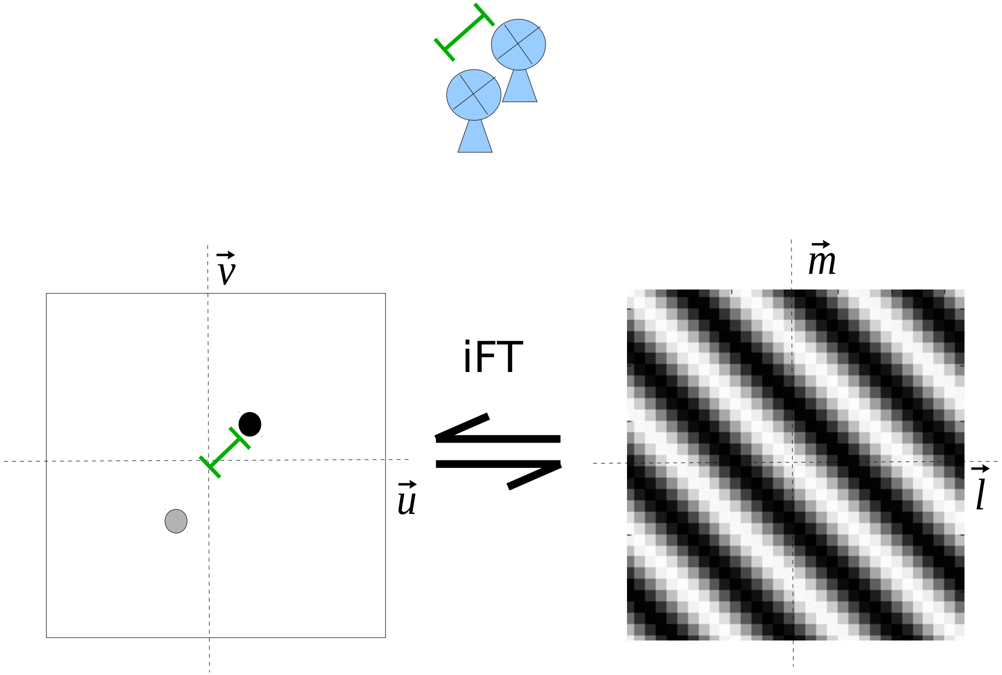
$\vec{u}, \vec{v}, \vec{b}$ (geometry)

Visibilities on the UV plane

Measure the spatial correlation of the E-field incident at each pair of antennas



N antennas
N(N-1)/2 antenna-pairs (baselines)

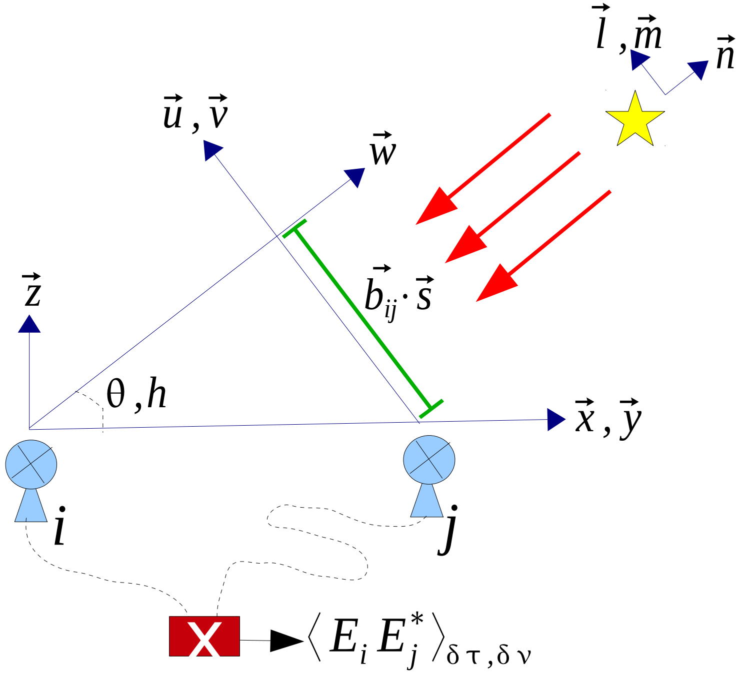


UV domain
(spatial frequency)

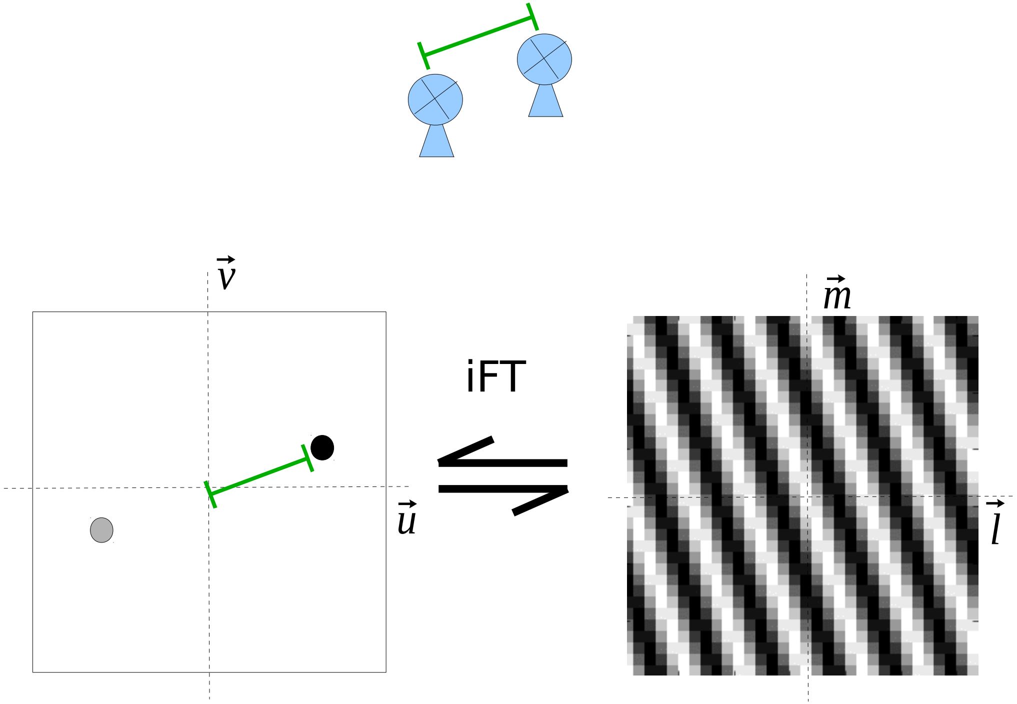
Image domain
(directions on sky)

Visibilities on the UV plane

Measure the spatial correlation of the E-field
incident at each pair of antennas



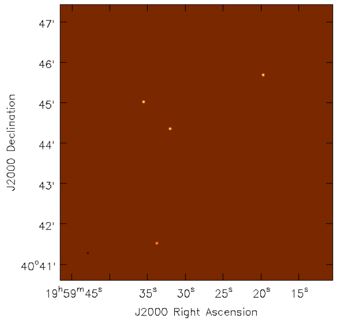
N antennas
N(N-1)/2 antenna-pairs (baselines)



UV domain
(spatial frequency)

Image domain
(directions on sky)

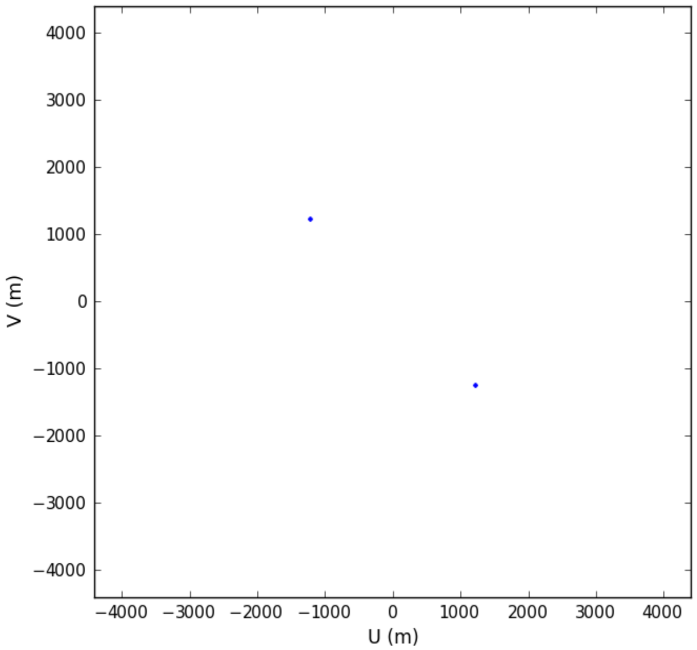
Spatial Frequency (uv) coverage + Observed Image



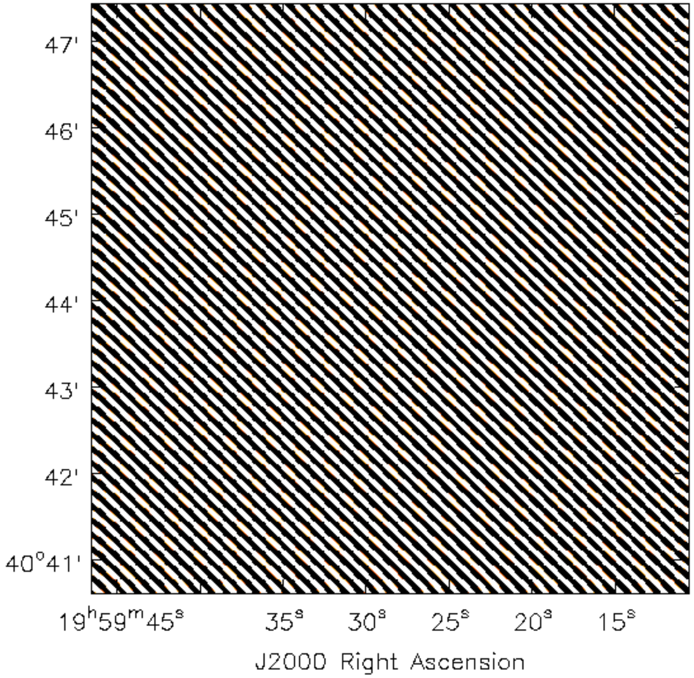
$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \frac{1}{\lambda} \begin{bmatrix} R(h,\theta) \end{bmatrix} \begin{bmatrix} \delta x \\ \delta y \\ \delta z \end{bmatrix}$$

Image of the sky
using **2** antennas

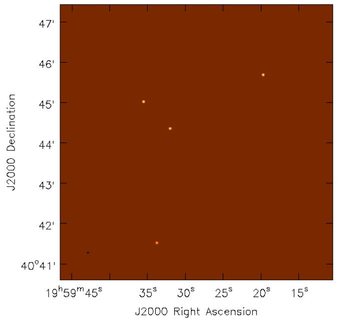
$$S(u,v)$$



$$I^{obs}(l,m)$$



Spatial Frequency (uv) coverage + Observed Image

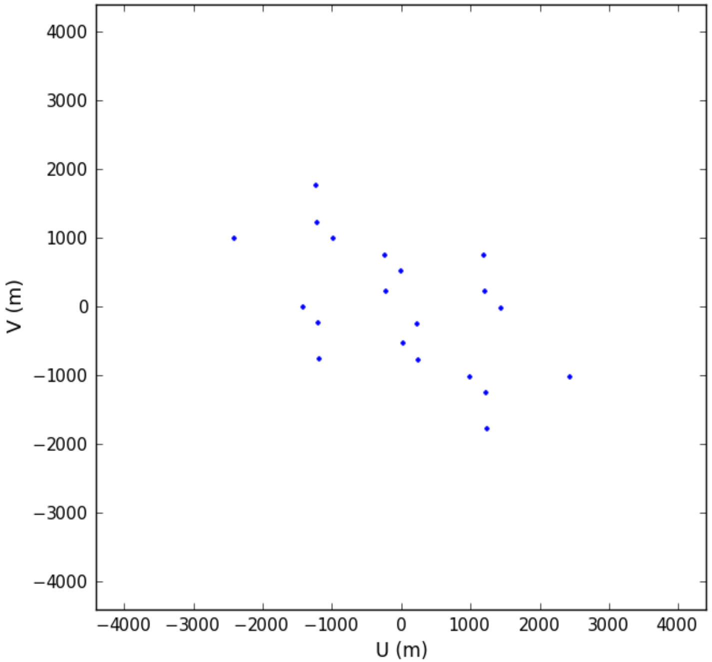


$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \frac{1}{\lambda} \begin{bmatrix} R(h,\theta) \end{bmatrix} \begin{bmatrix} \delta x \\ \delta y \\ \delta z \end{bmatrix}$$

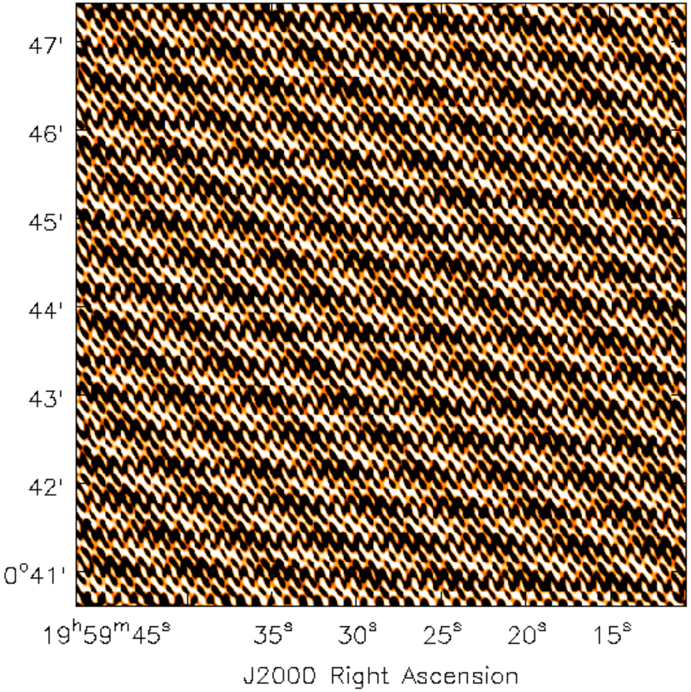
Image of the sky
using **5** antennas

"Aperture Synthesis"

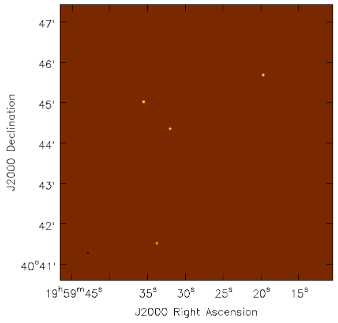
$S(u,v)$



$I^{obs}(l,m)$

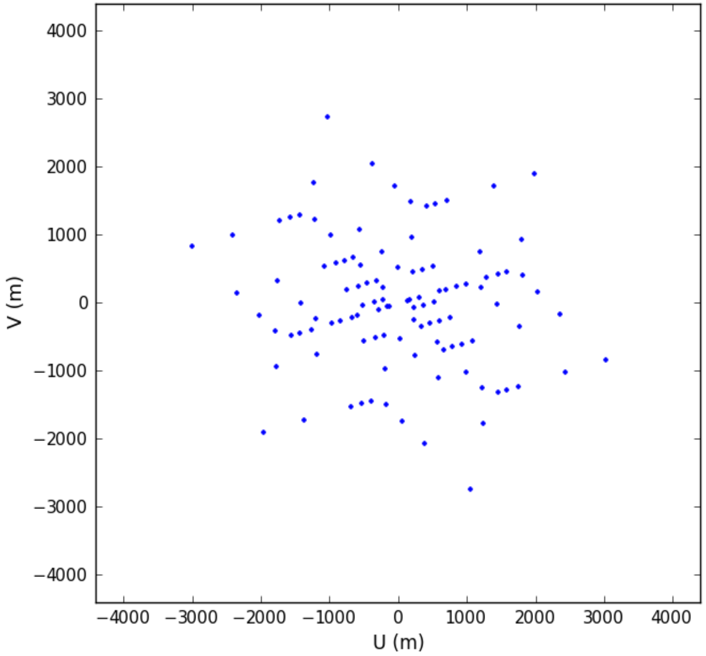


Spatial Frequency (uv) coverage + Observed Image



$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \frac{1}{\lambda} \begin{bmatrix} R(h,\theta) \end{bmatrix} \begin{bmatrix} \delta x \\ \delta y \\ \delta z \end{bmatrix}$$

$S(u,v)$



$I^{obs}(l,m)$

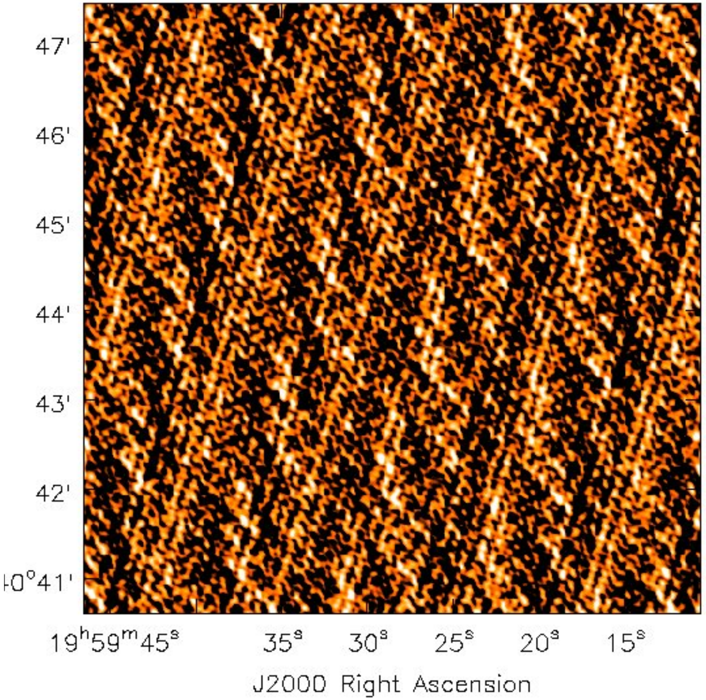
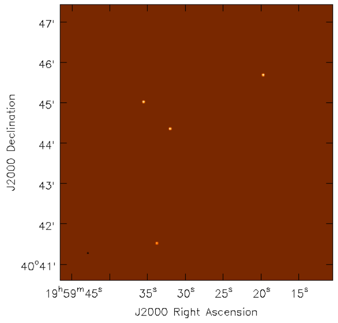


Image of the sky
using **11** antennas

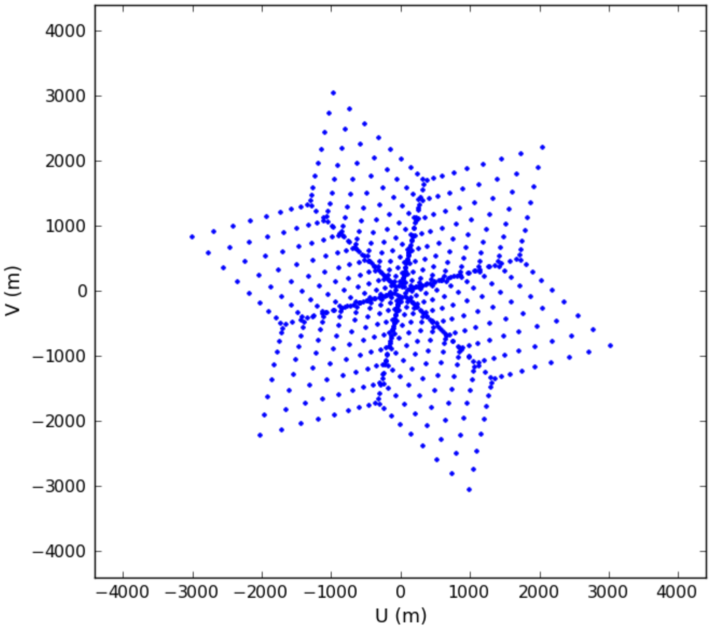
"Aperture Synthesis"

Spatial Frequency (uv) coverage + Observed Image



$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \frac{1}{\lambda} \begin{bmatrix} R(h,\theta) \end{bmatrix} \begin{bmatrix} \delta x \\ \delta y \\ \delta z \end{bmatrix}$$

$S(u,v)$



$I^{obs}(l,m)$

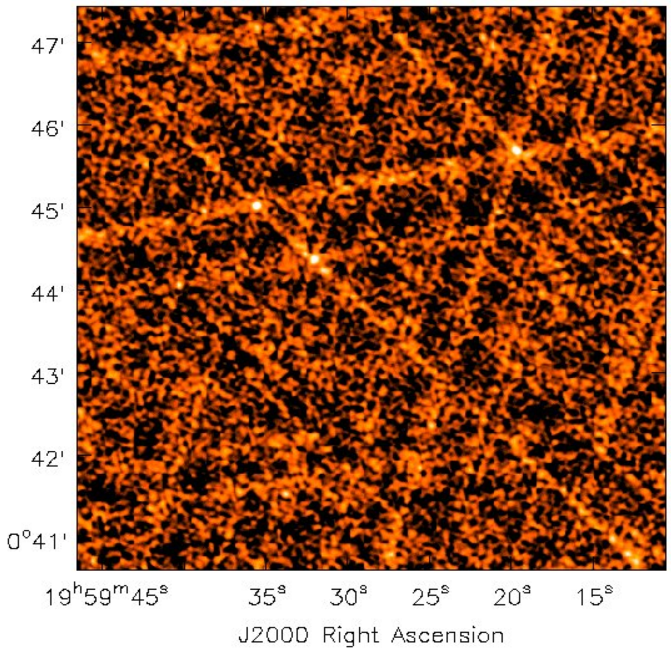
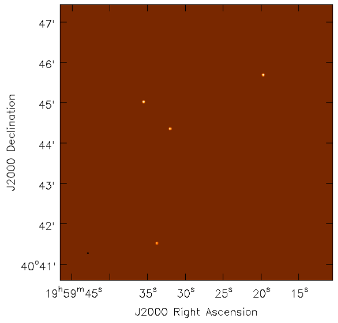


Image of the sky
using **27** antennas

"Aperture Synthesis"

Spatial Frequency (uv) coverage + Observed Image



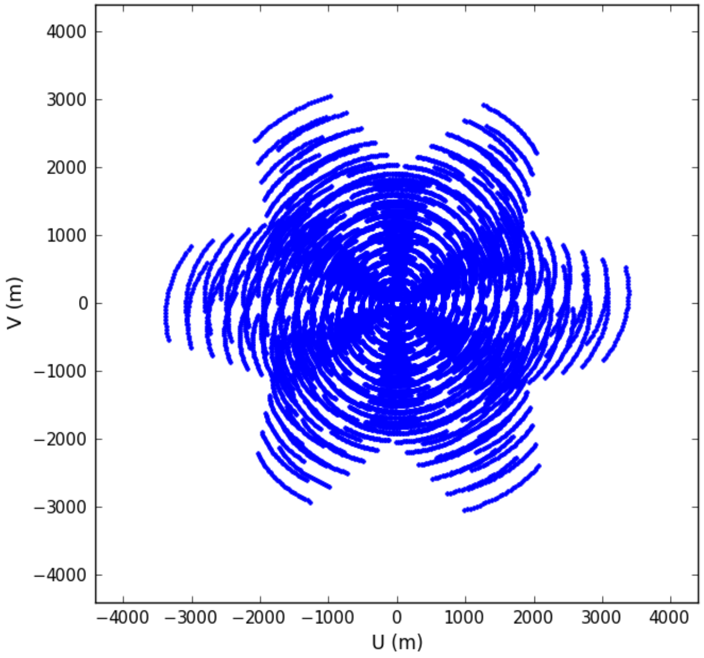
$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \frac{1}{\lambda} \begin{bmatrix} R(h,\theta) \end{bmatrix} \begin{bmatrix} \delta x \\ \delta y \\ \delta z \end{bmatrix}$$

Image of the sky
using 27 antennas

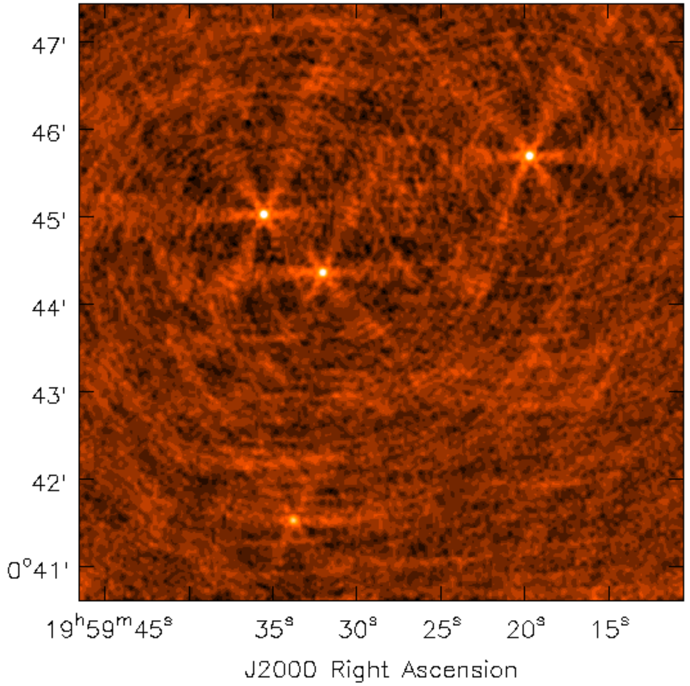
Observation : **2 hours**

"Earth Rotation Synthesis"

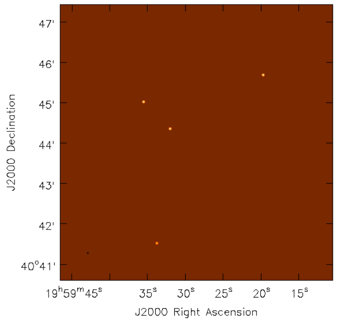
$S(u,v)$



$I^{obs}(l,m)$



Spatial Frequency (uv) coverage + Observed Image



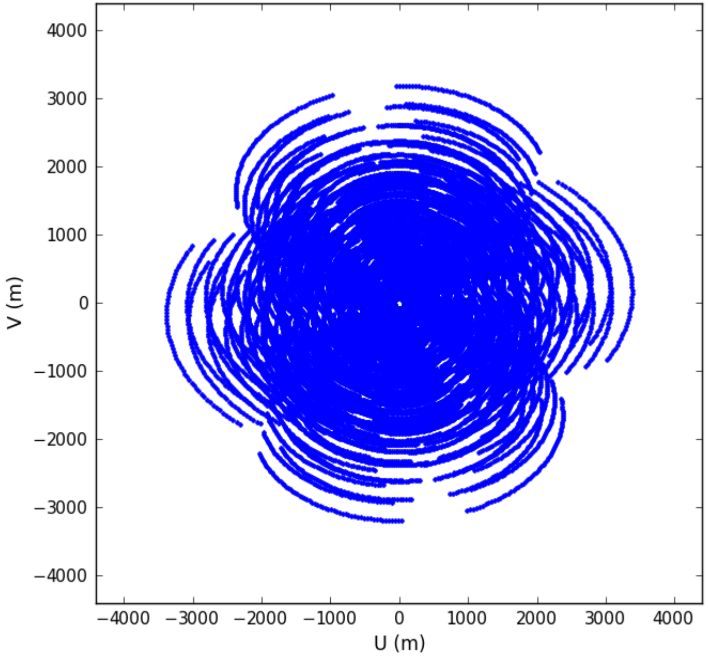
$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \frac{1}{\lambda} \begin{bmatrix} R(h,\theta) \end{bmatrix} \begin{bmatrix} \delta x \\ \delta y \\ \delta z \end{bmatrix}$$

Image of the sky
using 27 antennas

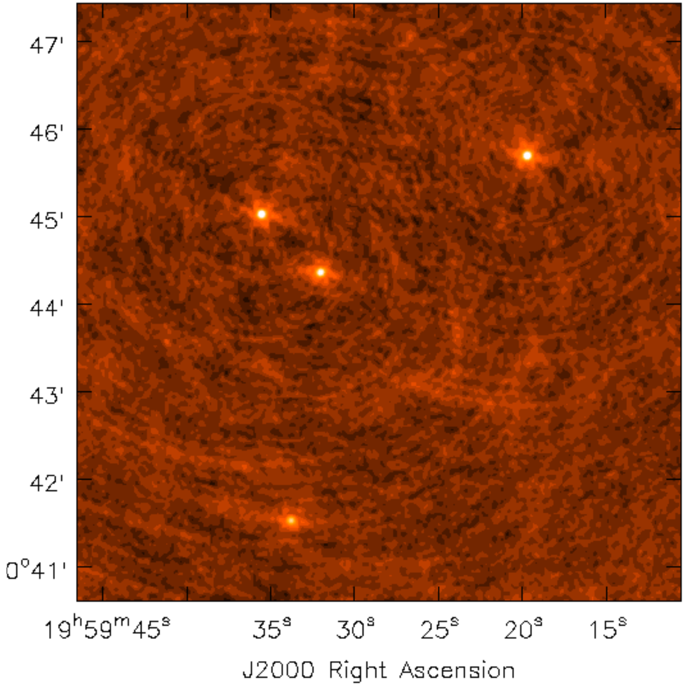
Observation : **4 hours**

"Earth Rotation Synthesis"

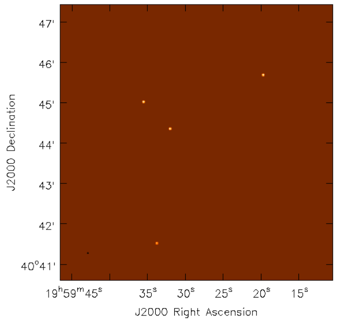
$S(u,v)$



$I^{obs}(l,m)$

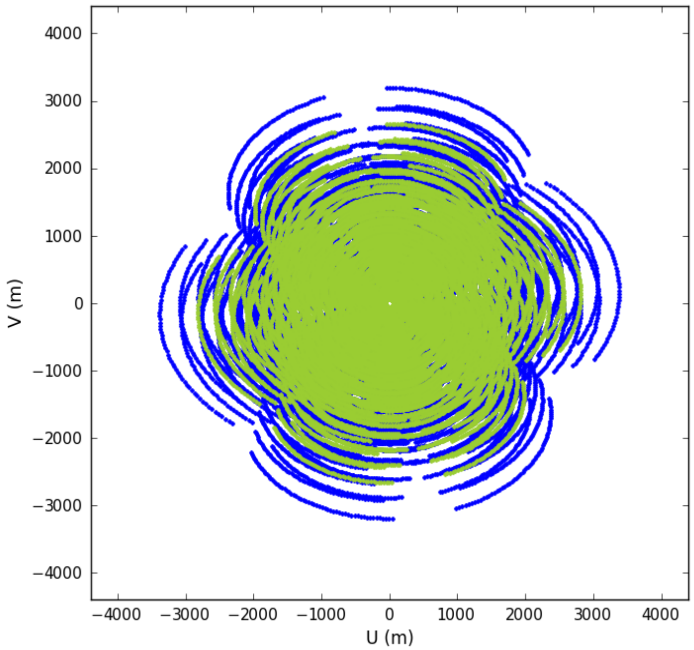


Spatial Frequency (uv) coverage + Observed Image



$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \frac{1}{\lambda} \left[R(h,\theta) \right] \begin{bmatrix} \delta x \\ \delta y \\ \delta z \end{bmatrix}$$

$S(u,v)$



$I^{obs}(l,m)$

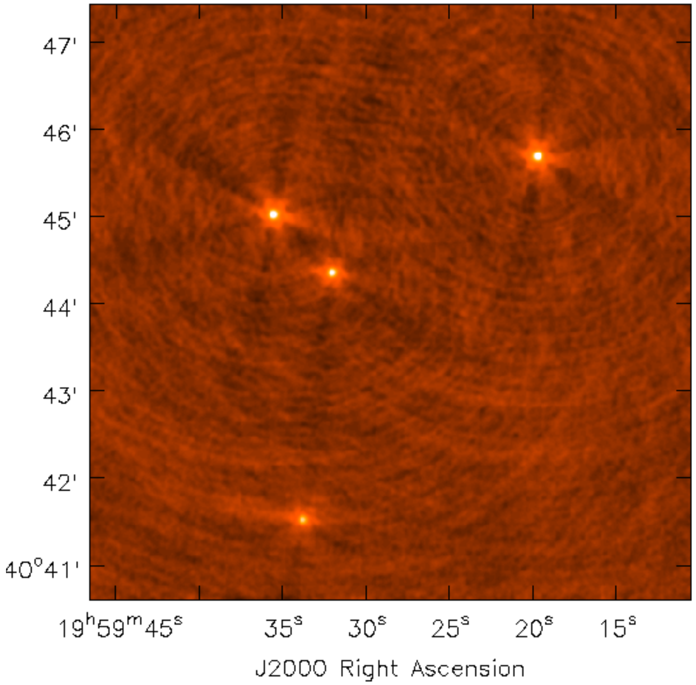
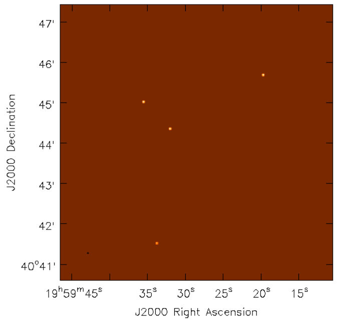


Image of the sky
using 27 antennas

Observation : 4 hours, **2 frequency channels**

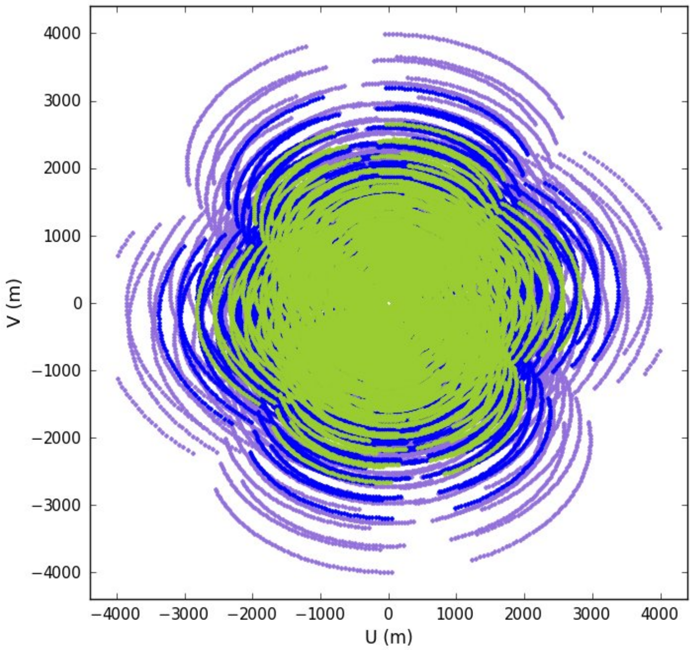
"Multi Frequency Synthesis"

Spatial Frequency (uv) coverage + Observed Image



$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \frac{1}{\lambda} \left[R(h,\theta) \right] \begin{bmatrix} \delta x \\ \delta y \\ \delta z \end{bmatrix}$$

$S(u,v)$



$I^{obs}(l,m)$

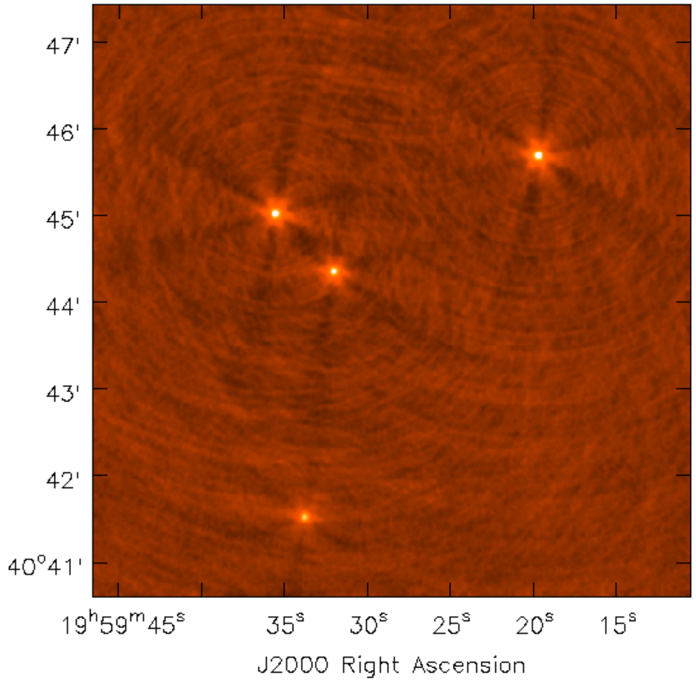
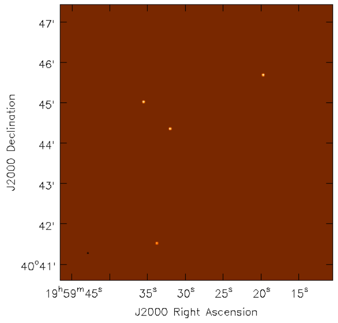


Image of the sky
using 27 antennas

Observation : 4 hours, **3 frequency channels**

"Multi Frequency Synthesis"

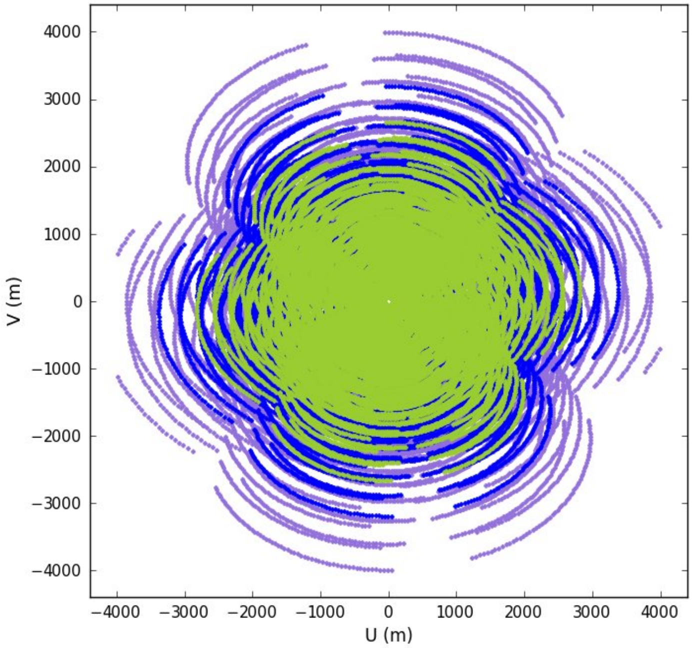
Spatial Frequency (uv) coverage + Observed Image



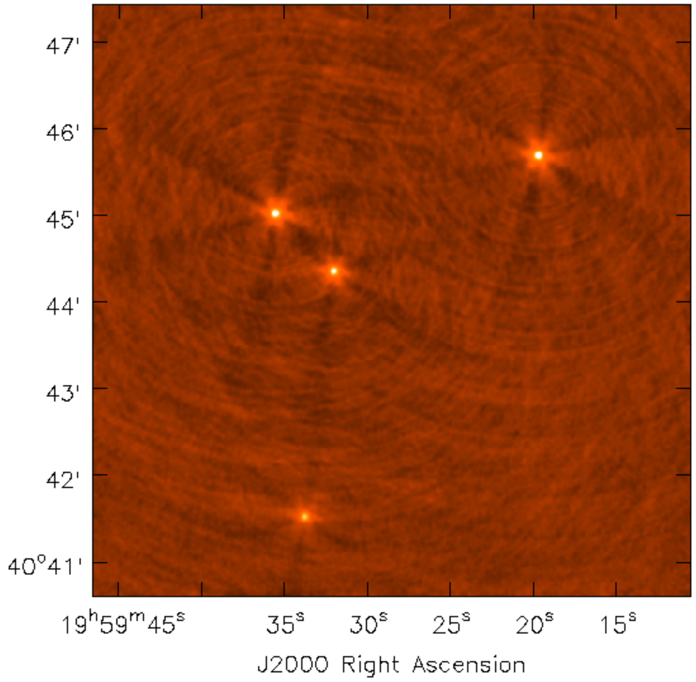
$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \frac{1}{\lambda} \begin{bmatrix} R(h,\theta) \end{bmatrix} \begin{bmatrix} \delta x \\ \delta y \\ \delta z \end{bmatrix}$$

Image of the sky
using 27 antennas

$S(u,v)$



$I^{obs}(l,m)$



Observation : 4 hours, 3 frequency channels

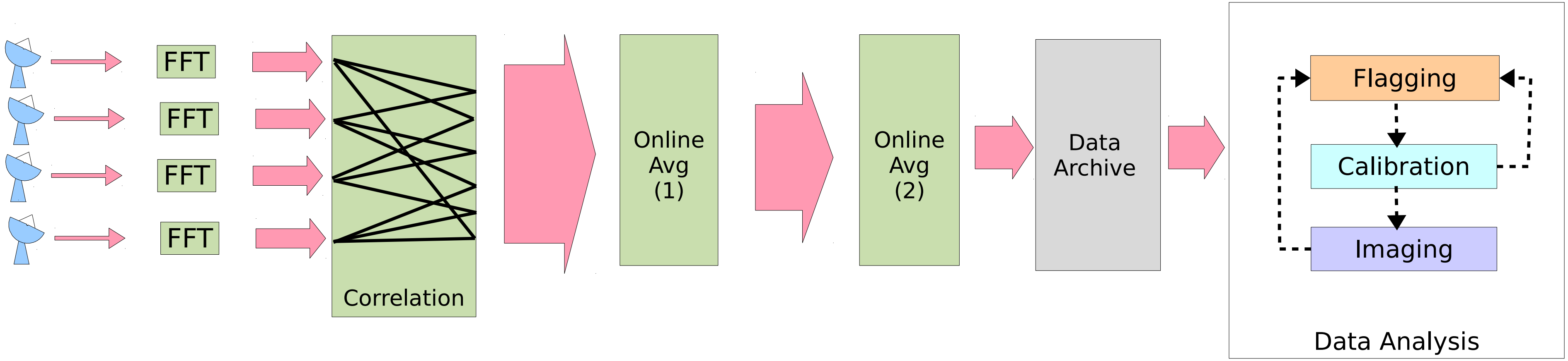
Point Spread Function

=> Imaging Properties

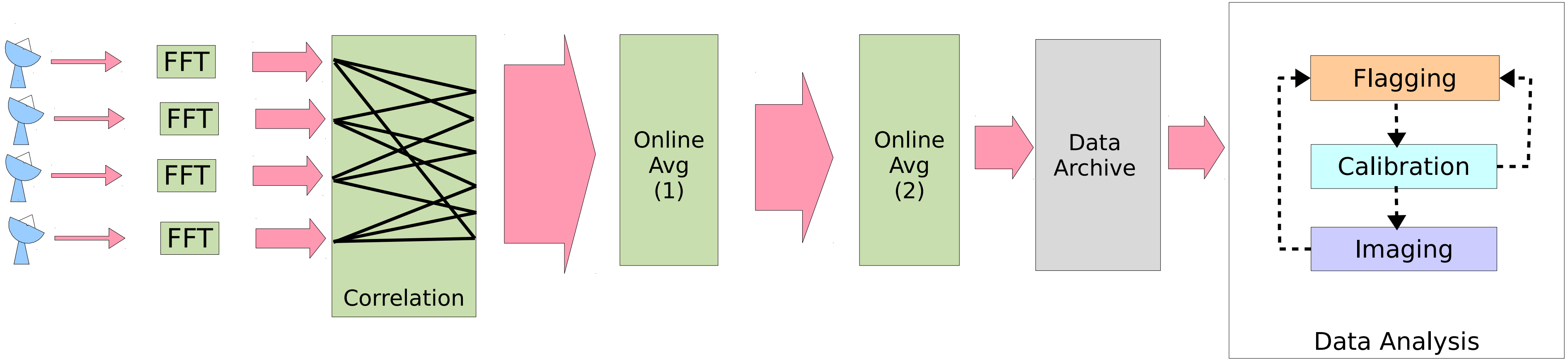
Outline

- Introduction to Radio Interferometry
- Data Management
  - Data Acquisition
  - Flagging, Calibration, Imaging
  - Pipelines and Automation
- Areas of HPC application and innovation

Data Acquisition and Analysis



Data Acquisition and Analysis

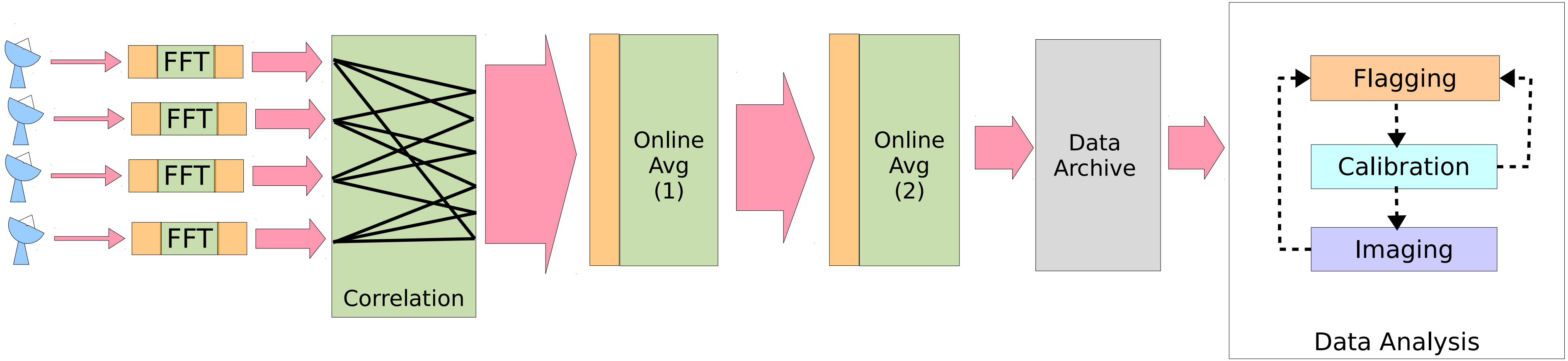


Real time   micro-sec   --> milli-sec   --> seconds

Digitize   Channelize   Multiply + Accumulate   Average   Average

Real Time System

Data Acquisition and Analysis



Real time   micro-sec   --> milli-sec   --> seconds

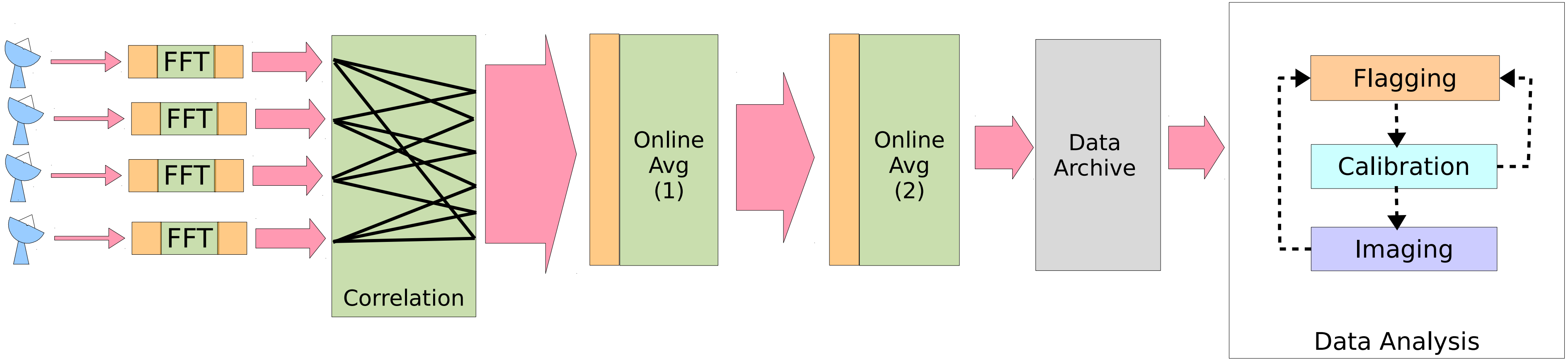
Digitize   Channelize   Multiply + Accumulate   Average   Average

+ *Outlier detection*

+ *Outlier detection*
+ *Real-time imaging*
Transient Detection

Real Time System

Data Acquisition and Analysis



Each observation is a database of N(N-1)/2 x N_time x N_chan x N_pol complex numbers

E.g. : N_time = 6 hours / 1 sec = 21600 timesteps
N_chan = 1 GHz / 1MHz = 1000 channels
N_pol = 4
N=27

=> [VLA : ~ 1 TB per day]

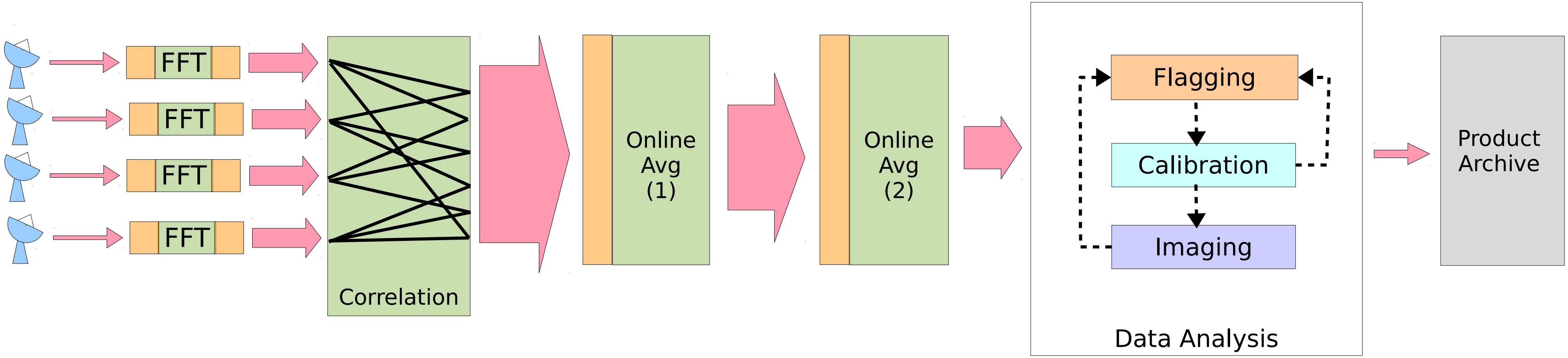
[ALMA (WB-upgrade) : ~ 100 TB per day]

[NgVLA : 100 TB to 1 PB per day]

[SKA : 4 TB/s into proc => ~ 100 PB per day]

Data Archive

Data Acquisition and Analysis



Processing Results are stored (for each observation) : N_xpix x N_ypix x N_chan x N_pol

– Image Cubes + Auxiliary information + Derived products

– Tools for image exploration

E.g. N_xpix, N_ypix : 1k → 20k
N_chan : 1 → 1k → 1M
N_pol : 4

Product Archive

Data Analysis

Flagging	Calibration	Imaging
Identify and mask corrupted data (RFI, Instrument errors, etc)	Derive and apply corrections to undo the effects of complex valued antenna gains	Reconstruct images by iterative model fitting while correcting for other instrumental effects

Flagging

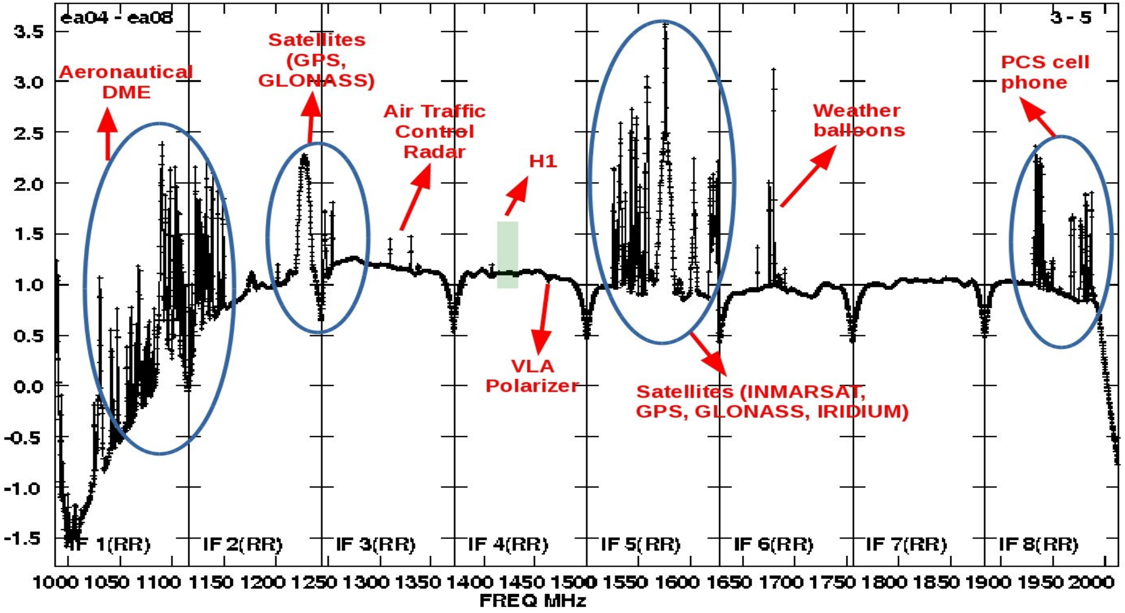
Flagging	Calibration	Imaging
Identify and mask corrupted data (RFI, Instrument errors, etc)	Derive and apply corrections to undo the effects of complex valued antenna gains	Reconstruct images by iterative model fitting while correcting for other instrumental effects

Identify and mask unusable data.

- Radio Frequency Interference
- Instrumental Errors & Effects

Algorithm :
- Outlier Detectors,
- Meta-data based flags/masks

Parallelism along multiple data dimensions



Calibration

Flagging

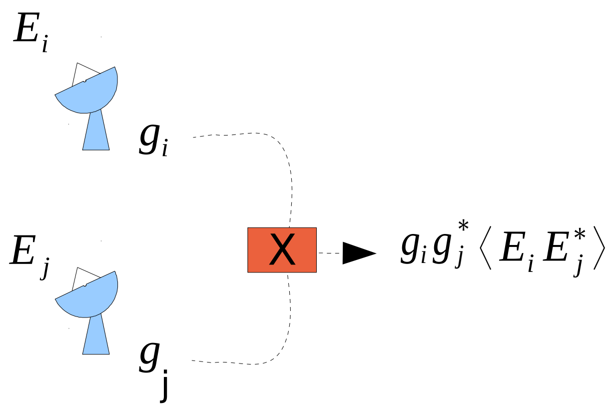
Identify and mask corrupted data (RFI, Instrument errors, etc)

Calibration

Derive and apply corrections to undo the effects of complex valued antenna gains

Imaging

Reconstruct images by iterative model fitting while correcting for other instrumental effects



- Observe a source where $\langle E_i E_j^* \rangle$ is known
- Use information from all ij to solve for g_i
- Divide out $g_i g_j^*$ from target data

Algorithms : Non-linear least squares solvers

Multi-stage process, each with different data views. Parallelism per stage across data dimensions

Imaging

Flagging

Identify and mask corrupted data (RFI, Instrument errors, etc)

Calibration

Derive and apply corrections to undo the effects of complex valued antenna gains

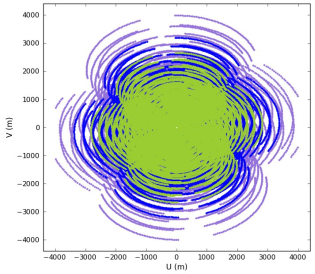
Imaging

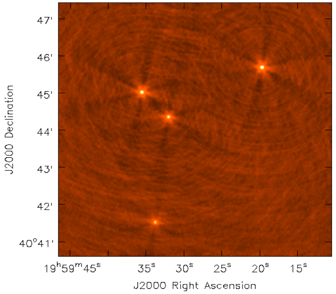
Reconstruct images by iterative model fitting while correcting for other instrumental effects

(1) Image Formation

- Place weighted measurements on a 2D grid

- Take iFFT





(2) Image Reconstruction

– Data : Incomplete sampling of the Fourier Space

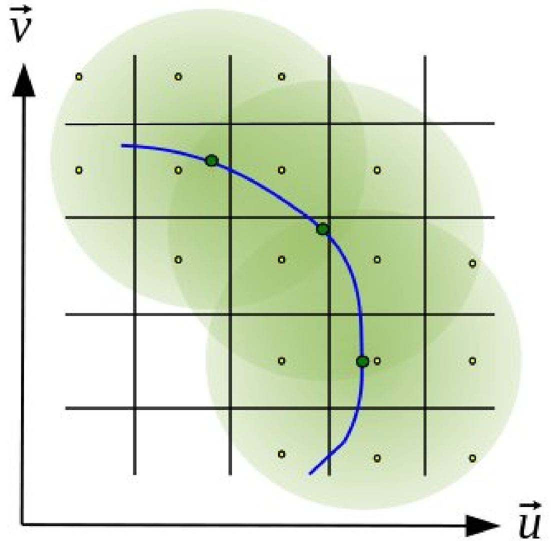
– Modeling : Iterative fitting to reconstruct a model of sky brightness

=> Remove the effect of the point-spread-function

Image Formation

Data : N_vis = N(N-1)/2 x N_time x N_chan x N_pol complex numbers

Gridding : Convolutional Resampling



Computing :

N_k x N_k : Footprint of convolution kernel

N_vis x N_k x N_k :

Multiplications + Additions

This is a compute hotspot

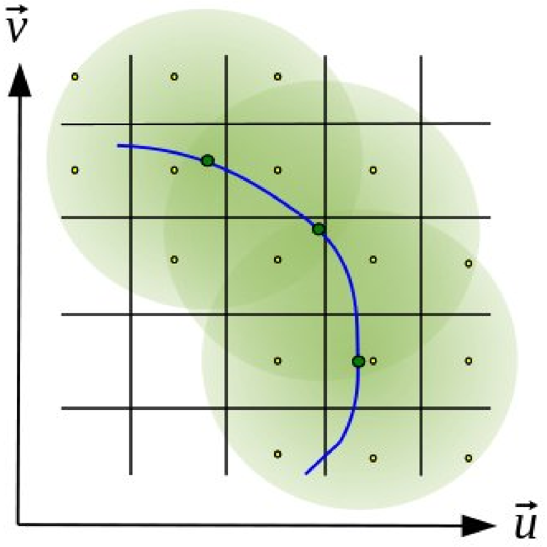
- Data parallelism
- GPUs (x100 speedup)

(Ref: ngVLA Computing Memo #5)

Image Formation

Data : N_vis = N(N-1)/2 x N_time x N_chan x N_pol complex numbers

Gridding : Convolutional Resampling



Computing :

N_k x N_k : Footprint of convolution kernel

N_vis x N_k x N_k :

Multiplications + Additions

This is a compute hotspot

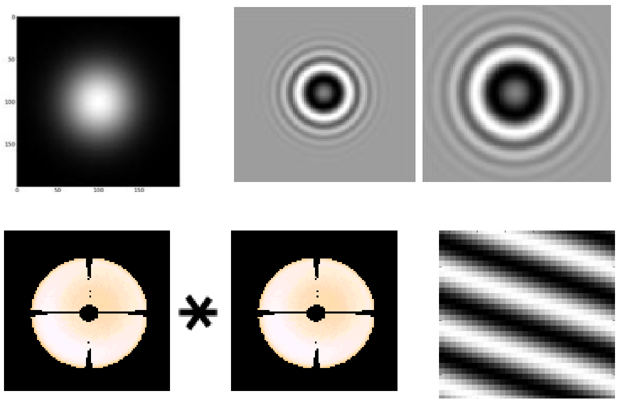
- Data parallelism
- GPUs (x100 speedup)

(Ref: ngVLA Computing Memo #5)

Types of Gridding Convolution Functions

- Depends on instrumental effects being corrected

- Range of N_k : 5 to few 100
  (runtime : 1hr → 10 days)



Iterative Image Reconstruction

The generalized forward problem $V^{obs} = [A] I^m + n$

The generalized inverse problem $I^m = [A]^{-1} V^{obs}$

L2 data regularization

\+ Sky model (multiscale, wideband, timevar)
\+ Solver/Optimizer with constraints/biases

Forward and Reverse transforms

Calc $\frac{\delta \chi^2}{\delta I^m}$

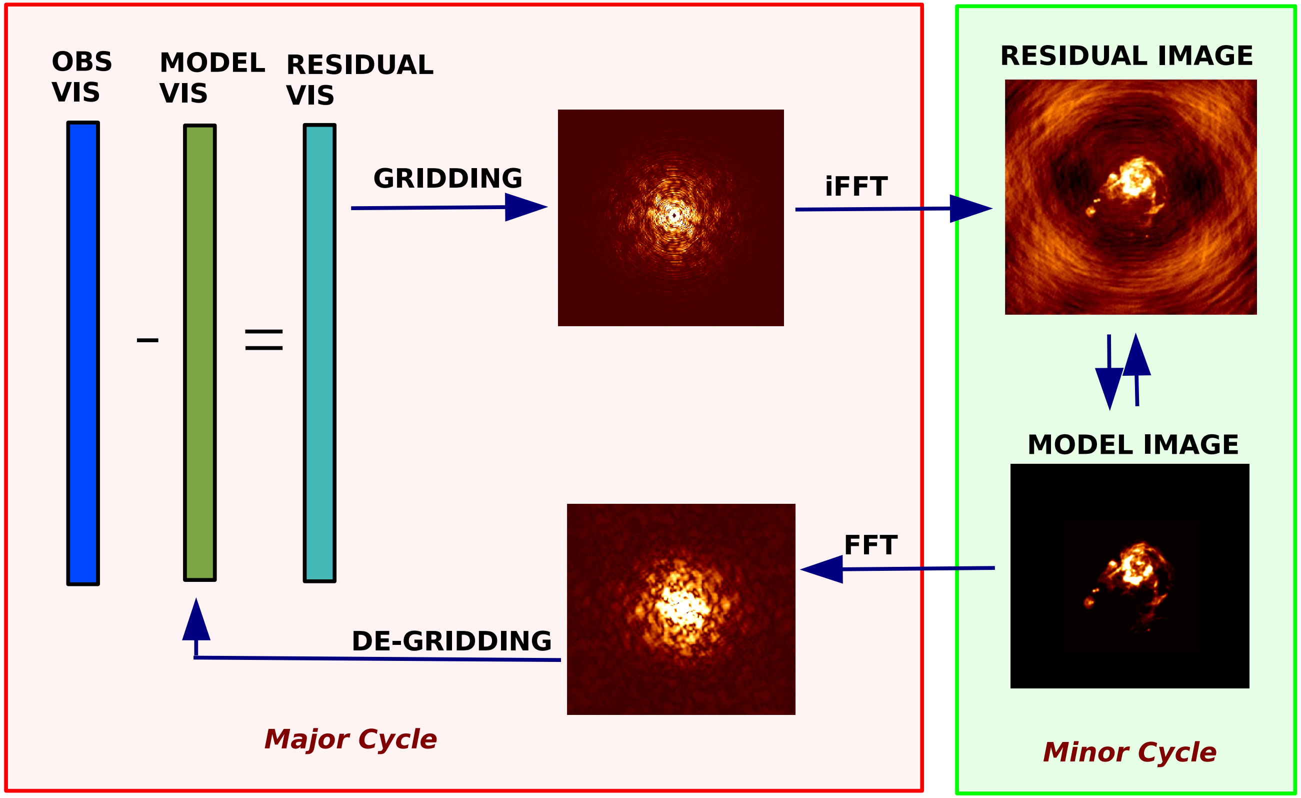


Image Reconstruction

Sky models
- Delta function
- Gaussians
- .....

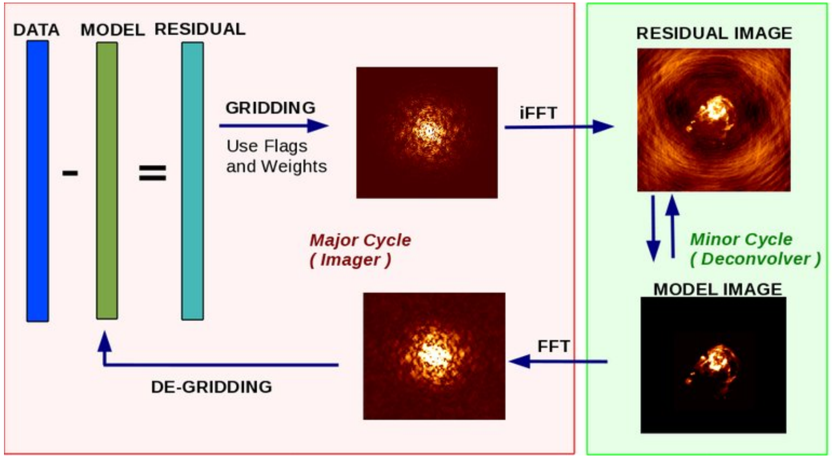
Algorithms
- Clean (greedy)
- Many other compressed sensing ideas

Imaging Compute Costs

Data I/O

Mostly Reads
Write once at the end.

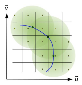
Partition in chunks by
“row” / “chan” / “time”



Imaging Compute Costs

Gridding : Convolutional resampling

Multi-threading
GPU acceleration

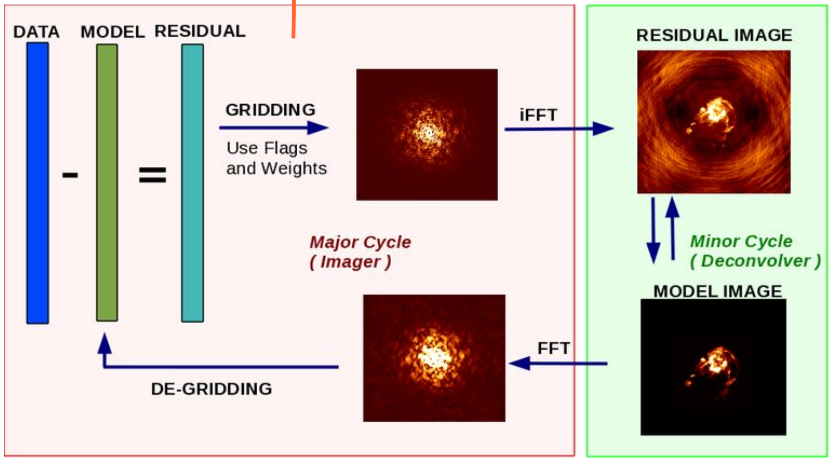


→ Adjust data ordering/access

Data I/O

Mostly Reads
Write once at the end.

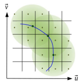
Partition in chunks by
“row” / “chan” / “time”



Imaging Compute Costs

Gridding : Convolutional resampling

Multi-threading
GPU acceleration

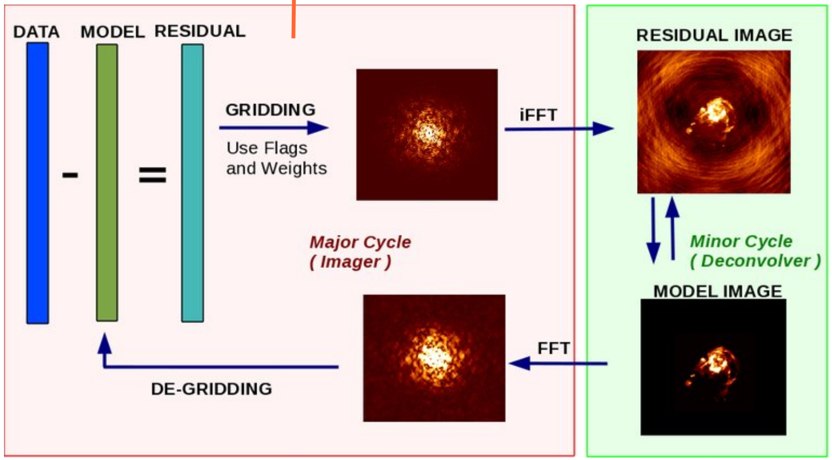


→ Adjust data ordering/access

Data I/O

Mostly Reads
Write once at the end.

Partition in chunks by
“row” / “chan” / “time”



Images : 4D cubes

FFTs, Math operations,
Fitting algorithms

Image reads/writes

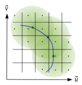
Multi-threading

Partitioning on “chan”
or “pixels”

Imaging Compute Costs

Gridding : Convolutional resampling

Multi-threading
GPU acceleration

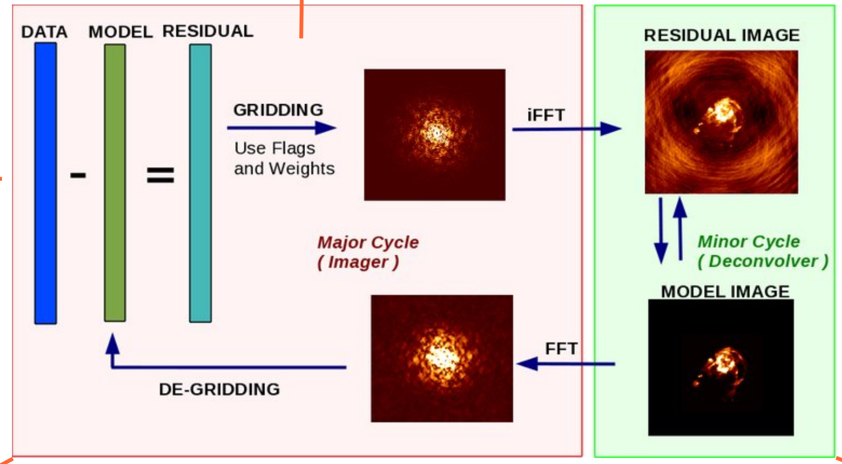


→ Adjust data ordering/access

Data I/O

Mostly Reads
Write once at the end.

Partition in chunks by
“row” / “chan” / “time”



Images : 4D cubes

FFTs, Math operations,
Fitting algorithms

Image reads/writes

Multi-threading

Partitioning on “chan”
or “pixels”

Number of iterations : 5 – 10 major cycle loops
100 to 10k minor cycle steps

Runtime varies by 1-2 orders of magnitude. Depends on data.

Science Ready Data Products : Data Analysis Pipelines

Flagging → Calibration → Imaging

Flagging

Outlier Detectors

Partition along "baseline"

or in chunks of "chan" / "time"

Calibration

Non-linear least squared solvers

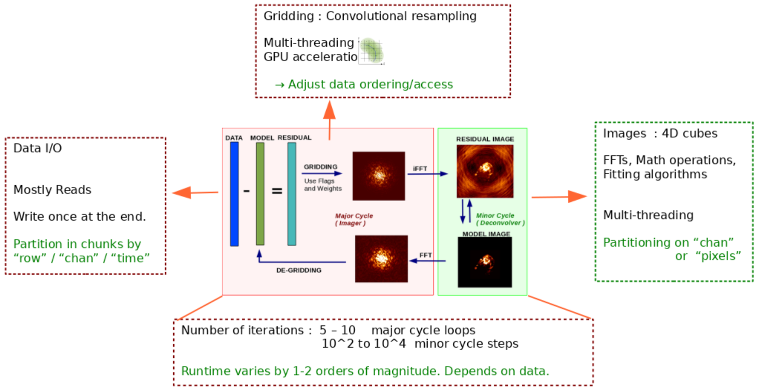
Partition along "time" or "chan" (but keep baselines together)

Multi-stage

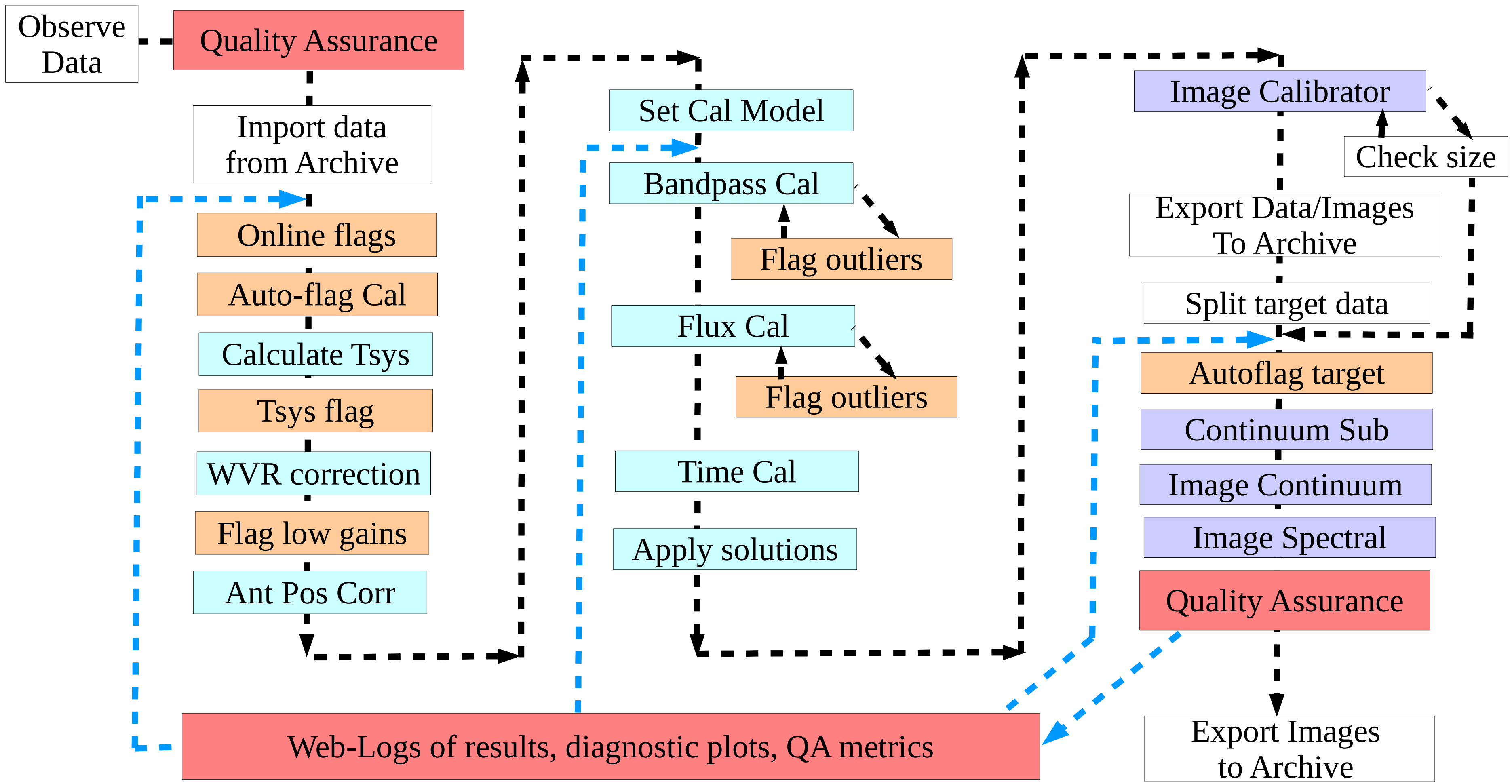
Data transforms (averaging, phase rotations, freq rebinning, etc) needed between calibration stages

Intermediate data shape changes

Imaging



Science Ready Data Products : Data Analysis Pipelines



Outline

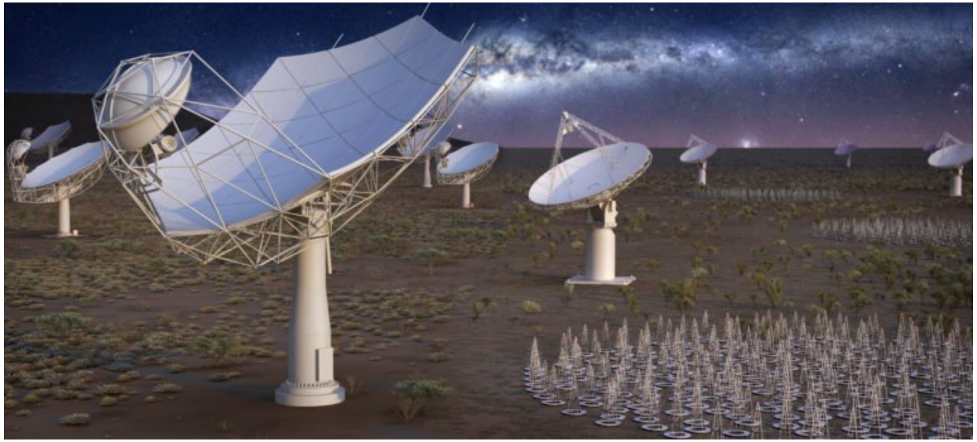
- Introduction to Radio Interferometry
- Data Management
  - Data Acquisition
  - Flagging, Calibration, Imaging
  - Pipelines and Automation
- Areas of HPC application and innovation

Going forward......

Data volumes will only increase (e.g. ngVLA, SKA....)

=> image noise reduces => instrumental effects easily seen => need complex algorithms

=> compute cost increases => manual intervention is harder => need HPC and automation



Square Kilometer Array (skatelescope.org)

2K dishes, 1M antennas , 50 MHz – 30 GHz

Next Generation VLA (ngvla.nrao.edu)

263 dishes (2 types) , 1-100 GHz

Pipeline Operations

Pipelines : A complex, data-dependent sequence of data processing steps

Computing : Optimize performance for the sequence of steps, not just each step on its own.

Types of projects :

– Surveys : Homogeneous observational setup and analysis steps.
– Targets : Diverse setups and analysis strategies. Need to support experimentation

Observatory Operations :

– Run pipelines for multiple datasets, optimizing for throughput. Keep up with observing rate.

References :

- SKA Science Data Processor : http://ska-sdp.org/publications/sdp-cdr-closeout-documentation
- ngVLA size of computing (imaging) : https://library.nrao.edu/public/memos/ngvla/NGVLAC_04.pdf

The R&D frontier

Algorithms :

- *Flagging* : Strategies targeted to different types of RFI, spectrum sharing, etc...
- *Calibration* : Wide-field and direction dependent solutions
- *Imaging* : A variety of sky models, instrument models, objective functions and regularizers, optimization strategies, the use of prior knowledge

Computing :

- Parallelization of data and algorithms
- GPUs for compute hotspots
- Scaleable compute frameworks (e.g. dask...)
- System design : Managing complexity

Automation :

– Heuristics and ML/AI for pattern classification, data inspection and decision automation, telescope monitoring and control with feedback, image and spectrum science analysis, etc.

Most radio astronomy observatories are engaged in these activities (with partners)