Processing MeerKAT data at IDIA*

IDIA Pipelines

Bradley Frank (SARAO / IDIA)
David Aikema, Jordan Collier, Srikrishna Sekhar,
Russ Taylor

***FOR MIGHTEE**

support@ilifu.ac.za

IDIA SKA AFRICA SQUARE KILOMETRE ARRAY NRF National Research Foundation

Pipeline Talks

- Talk 1: This Talk — High Level Overview of IDIA Pipelines.
- Talk 2: Jordan Collier — Usage.
- Talk 3: Srikrishna Sekhar — Under the Hood.

IDIA

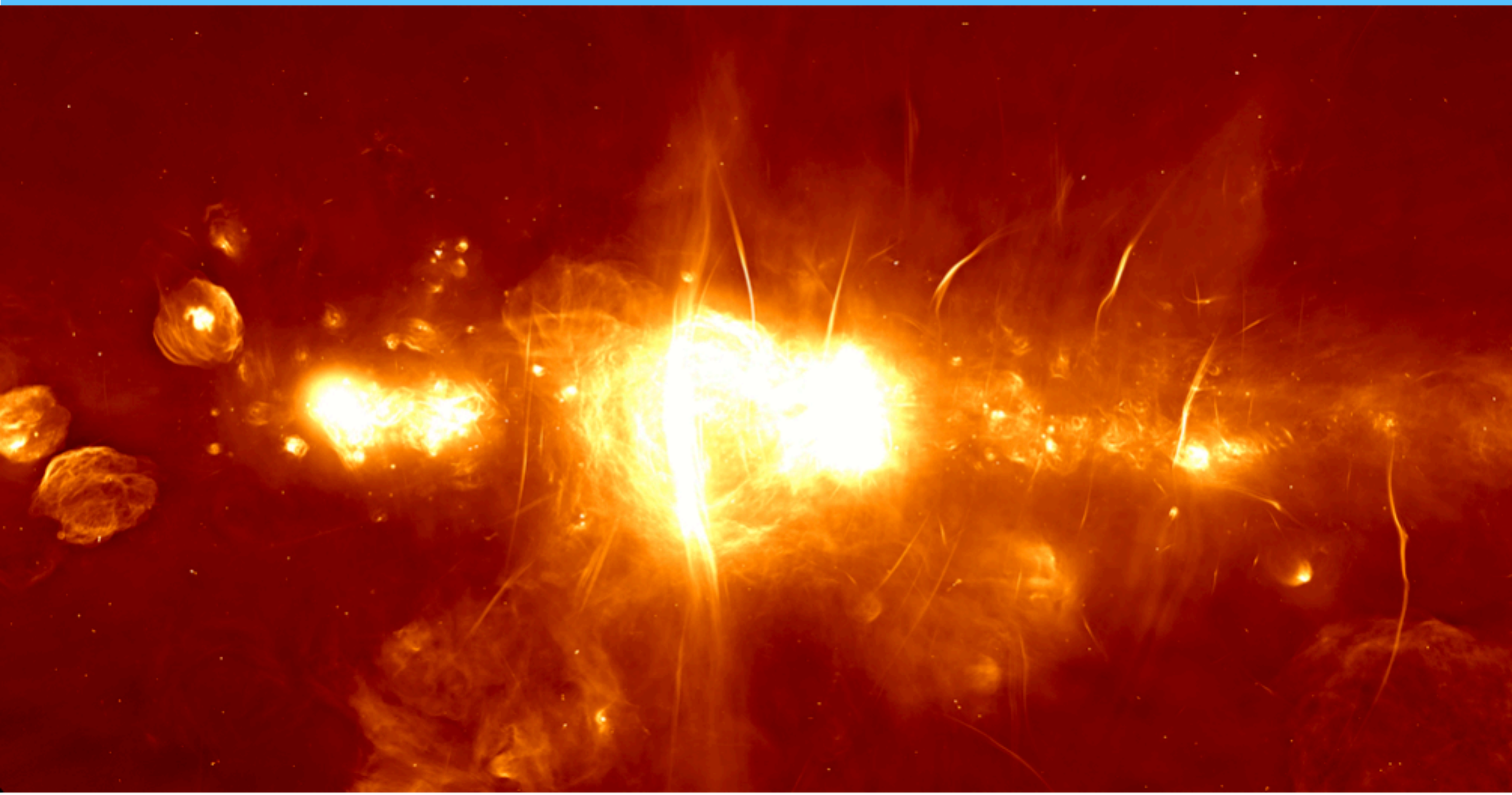
- The Inter-University Institute for Data Intensive Astronomy
  - OpenStack Cloud.
  - University of Cape Town
  - University of the Western Cape
  - University of Pretoria
  - Sol Plaatjie University
- Integrated into Tier 2 Data Intensive Research Facility:
  - ILIFU — IDIA and CBIO (Bioinformatics).

IDIA

- ***Pathfinder*** **Science Regional Data Centre.**
- A Research Service
  - Data Processing, Analysis, Storage, Access, Transport.
  - Support transparent, effective, reproducible science.
- Any member of a MeerKAT LSP can get access — provided they have permission from PIs.
  - http://users.idia.ac.za/access/
  - IDIA Involvement: **Most** if not **ALL** MeerKAT/M16/AR1.5 Science products!
- **LSP Feedback: bradley@idia.ac.za / bfrank@ska.ac.za / support@idia.ac.za**

IDIA SKA AFRICA SQUARE KILOMETRE ARRAY NRF National Research Foundation

MeerKAT



Radio Continuum Mosaic: Galactic Centre, Ian Heywood, SARAO (Processed at IDIA)

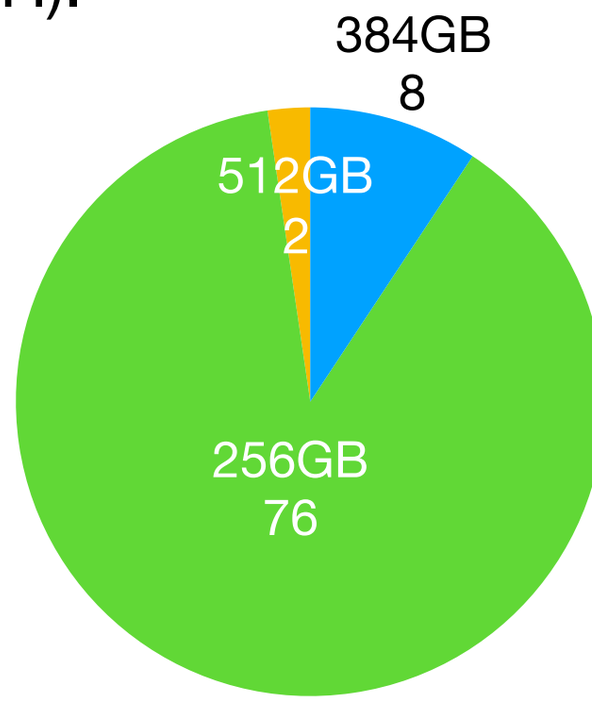
IDIA

SKA AFRICA
SQUARE KILOMETRE ARRAY

NRF National Research Foundation

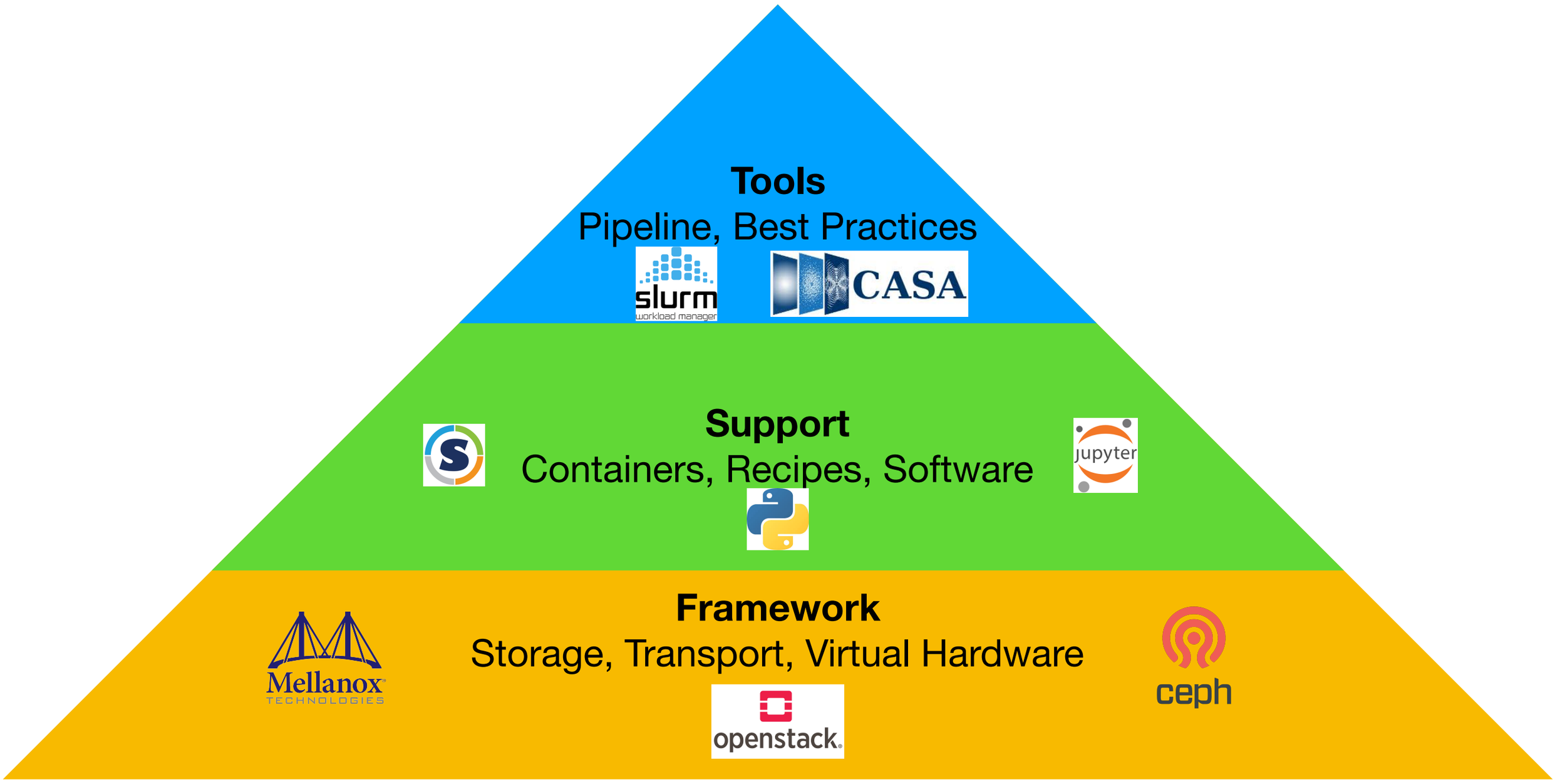
IDIA Kit

- Current integration into ILIFU.
  - ILIFU: Joint Cloud Centre for Astronomy and Bioinformatics.
  - Compute
    - 127 Nodes, 16-24 cores/node @ 2.6GHz (hyper-threading).
    - VHPC: 50 Nodes (256G/32VCPU) currently available (CEPH).
    - ~ 6 Fat Nodes (256G/32VCPU) — BeeGFS.
- Storage
  - 19 Nodes, 3.3PB raw (~1PB available to Pipelines/LSP).
  - BeeGFS (OS) / CEPH (LT).



IDIA SKA AFRICA SQUARE KILOMETRE ARRAY NRF National Research Foundation

IDIA Pipelines



IDIA
SKA AFRICA
SQUARE KILOMETRE ARRAY
NRF
National Research Foundation

General User Model

Control

Head Node

Control

Control

Fat Node

Worker Nodes

High-Speed
Network Mount

Data

Distributed Storage
BeeGFS / CEPH

Software: Singularity Containers
Workflow/Resource on VHPC:
SLURM

racetrack.idia.ac.za

File Edit View Run Kernel Hub Tabs Settings Help

ic5063 › 951to1100

Name	Last Modified
2000to2500_ap3	13 days ago
2000to2500.pygmask	13 days ago
951to1100_t.mms	13 days ago
951to1100_t.mms.0.0.alpha	13 days ago
951to1100_t.mms.0.0.alpha.error	13 days ago
951to1100_t.mms.0.0.gridwt_moswt	13 days ago
951to1100_t.mms.0.0.image.tt0	13 days ago
951to1100_t.mms.0.0.image.tt1	13 days ago
951to1100_t.mms.0.0.mask	13 days ago
951to1100_t.mms.0.0.model.tt0	13 days ago
951to1100_t.mms.0.0.model.tt1	13 days ago
951to1100_t.mms.0.0.pb.tt0	13 days ago
951to1100_t.mms.0.0.psf.tt0	13 days ago
951to1100_t.mms.0.0.psf.tt1	13 days ago

Launcher

data/ic5063/951to1100

Notebook

Python 3

Console

ipynbs — IPython: frank/laduma — -bash

~/work/idia-pipelines.github.io — IPython: Users/frank — -bash

~/work/ipynbs — jupyter-notebook ▸ python

```
frank@ilifu-slurm-login:~$ sinfo
PARTITION         AVAIL  TIMELIMIT  NODES  STATE NODELIST
Main*                up 3-00:00:00     20  down* slwrk-[010-012,014-019,021-022,024-025,03
Main*                up 3-00:00:00     11    mix slwrk-[006-007,009,013,020,026-028,030,03
Main*                up 3-00:00:00      4  alloc slwrk-[008,023,029,033]
Test01               up 3-00:00:00      8   idle slwrk-[041-048]
JupyerSpawnerONLY    up   infinite      2    mix slwrk-[001-002]
JupyerSpawnerONLY    up   infinite      3   idle slwrk-[003-005]
frank@ilifu-slurm-login:~$ sq
frank@ilifu-slurm-login:~$ squeue
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
           1084185      Main quick_tc    frank PD       0:00      5 (Dependency)
           1084178      Main cal_xx_y    frank PD       0:00      5 (Dependency)
           1084179      Main flag_rou    frank PD       0:00      5 (Dependency)
           1084180      Main run_setj    frank PD       0:00      5 (Dependency)
           1084182      Main cal_xy_y    frank PD       0:00      5 (Dependency)
           1084183      Main    split    frank PD       0:00      5 (Dependency)
           1084200      Main quick_tc    frank PD       0:00      3 (Dependency)
           1084252      Main quick_tc jcollier PD       0:00      4 (Dependency)
           1084243      Main run_setj jcollier PD       0:00      4 (Dependency)
           1084245      Main cal_xx_y jcollier PD       0:00      4 (Dependency)
           1084246      Main flag_rou jcollier PD       0:00      4 (Dependency)
           1084247      Main run_setj jcollier PD       0:00      4 (Dependency)
           1084249      Main cal_xy_y jcollier PD       0:00      4 (Dependency)
           1084250      Main    split jcollier PD       0:00      4 (Dependency)
           1084194      Main flag_rou    frank PD       0:00      3 (Dependency)
           1084195      Main run_setj    frank PD       0:00      3 (Dependency)
           1084197      Main cal_xy_y    frank PD       0:00      3 (Dependency)
           1084198      Main    split    frank PD       0:00      3 (Dependency)
           1083814      Main  wsclean   bright PD       0:00      1 (Dependency)
           1083829      Main  wsclean   bright PD       0:00      1 (Dependency)
           1083830      Main     flag   bright PD       0:00      1 (Dependency)
           1083831      Main phase_ca   bright PD       0:00      1 (Dependency)
           1083832      Main  wsclean   bright PD       0:00      1 (Dependency)
           1083833      Main     flag   bright PD       0:00      1 (Dependency)
           1083834      Main phase_ca   bright PD       0:00      1 (Dependency)
           1083835      Main  wsclean   bright PD       0:00      1 (Dependency)
           1084289      Main phase_ca   bright PD       0:00      1 (Dependency)
           1084290      Main  wsclean   bright PD       0:00      1 (Dependency)
1084146_[10-81%10]      Main pol_cube   krishna PD       0:00      1 (JobArrayTaskLimit)
```

File View Layout Help

X Profile: Cursor

Y Profile: Cursor

Z Profile: Cursor

Declination

Right ascension

X coordinate

Y coordinate

Frequency (GHz)

Value (Jy/beam)

Render Configuration

Log

90% 95% 99% 99.5% 99.9% 99.95% 99.99% 100% Custom

Histogram Per-Channel

Scaling Linear

Color map

Min -0.0008533514

Max 0.0047087299

Animator

Frame rate

Channel

Count

Data Transfer
Current State

- IDIA/SARAO Data Transfer Node.
  - Raw data scraped off S3 database (Rados + NPY Array).
  - Converted to MS -> DTN.
  - DTN *push* request initiated.
  - GridFTP Transfer queued (managed by FTS).
  - Received at IDIA cluster.

Data Transfer
In Development

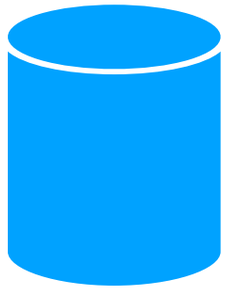
- PI checks archive interface (via VPN) for data.
- IDIA affiliated PI can *Push-To-IDIA* (its a button).
- MS data is transferred directly to appropriate IDIA directory.
- Transfer progress can be monitored on archive dashboard.

MeerKAT

Correlator SDP Cal (L0) (Site/Karoo)

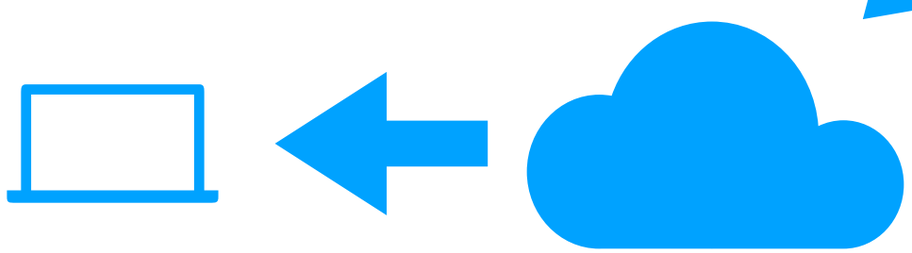
100Gb/s

Archive (L0+L1) (CHPC/Cape Town)



Data Transfer Node(s) (CHPC/Cape Town)

10/40/100 Gb/s



IDIA*

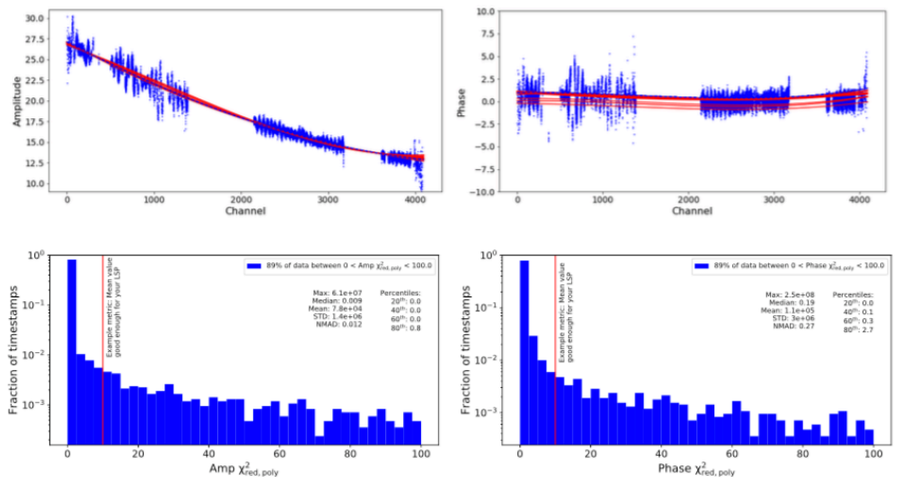
Data Quality Assurance

Table 1: Selected requirements and specifications from various MeerKAT LSPs

LSP	Sensitivity (μJy beam^{-1})	Dynamic Range	Velocity Resolution (km s^{-1})	Redshift Range	Area (deg^2)	RM precision (rad m^{-2})	Column Density Sensitivity	H I Mass ($M_\odot$)
MIGHTEE (L-band)	2^a / 90^b	$\geq 10^{5c}$	6^d	$z \lesssim 0.5$	20	$\sim$1	$\sim 1\ M_\odot\ pc^{-2^e}$	$\gtrsim 2 \times 10^{9f}$
MIGHTEE (S-band)	1^g	$\geq 10^{5c}$	-	$z \lesssim 0.5$	5.5	$\sim$1	-	-
MIGHTEE (UHF-band)	6^a	$\geq 10^{5c}$	-	$z \lesssim 0.5$	3.5	$\sim$1	-	-
LADUMA (L-band)	45^b	-	6^d	$0 \leq z \leq 0.58$	0.9–2.2	-	-	$\sim 10^{7.5-10.5}$
LADUMA (UHF-band)	26^h	-	8^i	$0.42 \leq z \leq 1.45$	1.8–5.4	-	-	$\sim 10^{9.2-10.5}$
FORNAX	100^j	-	$\sim$1	$z \sim 0$	$\sim$12	-	$\sim 0.1^k - 5^l \times 10^{19}$cm^{-2}	$\gtrsim 5 \times 10^{5l}$
MHONGOOSE	74^j	-	16	$z \sim 0$	$\sim 45^m$	$\lesssim 1$	$0.55^n - 7.5^o \times 10^{18}$cm^{-2}	$\sim 10^{6-11}$

Table 2: Data quality metrics for a simple calibration pipeline. This table is not exhaustive, but represents a selection of metrics we have drafted.

Step	Distribution / plot	Statistic / Metric	Computation	Tolerance
Bandpass calibration	Calibrated bandpass (amplitude and phase) as a function of frequency (per timestamp)	Residual of polynomial fit	Mean reduced χ squared of polynomial fit to each timestamp, compared to reference timestamp[a]	< 10
	Calibrated bandpass (amplitude and phase) as a function of frequency	Normalised median absolute deviation	Maximum normalised absolute deviation from polynomial fit	< 5
Bandpass flagging	Calibrated amplitude as a function of frequency	Fraction of channels flagged	Number of channels with >50% of visibilities flagged divided by total number of channels	< 1
Phase calibration	Complex gain solutions as a function of time	Outlier metric	Running median?	?
	Calibrated amplitude as a function of frequency	Normalised median absolute deviation	Normalised median absolute deviation from polynomial fit	< 5
Phase cal. flagging	Calibrated amplitude as a function of time	Fraction of timestamps flagged	Number of timestamps with >50% of visibilities flagged divided by total number of timestamps	< 1
Target flagging	Amplitude as a function of frequency	Fraction of data flagged	Fraction of visibilities flagged divided by total visibilities	$< 20\%$
	Amplitude as a function of frequency	Normalised median absolute deviation	Normalised median absolute deviation	< 5



IDIA SKA AFRICA SQUARE KILOMETRE ARRAY NRF National Research Foundation

IDIA Pipelines

- ***processMeerKAT*** Pipeline.
  - **PARALLELISED** Package for processing on HPC (SLURM + ILIFU Cluster).
    - Flagging (RFLAG/TFCROP), Full Stokes Pol, Applycal, Quick Imaging.
    - Generalised for use on PBS/Torque controlled system.
  - **Robust**, *generic*, fast implementation of *a'priori* calibration (including flagging).
    - Easy to use, transparent, reproducible.
  - General purpose Selfcal.
  - Aim: T(cal) ~ T(obs)
- **Framework:** Best practices on how to use SLURM and MPICASA.
- **Developer's Guide:** How to write and include your own modules in the pipeline.

IDIA Pipelines

Welcome to the IDIA Pipelines Website. Here you'll find help and documentation to help you do calibration and imaging of Radio Data Using the IDIA system.

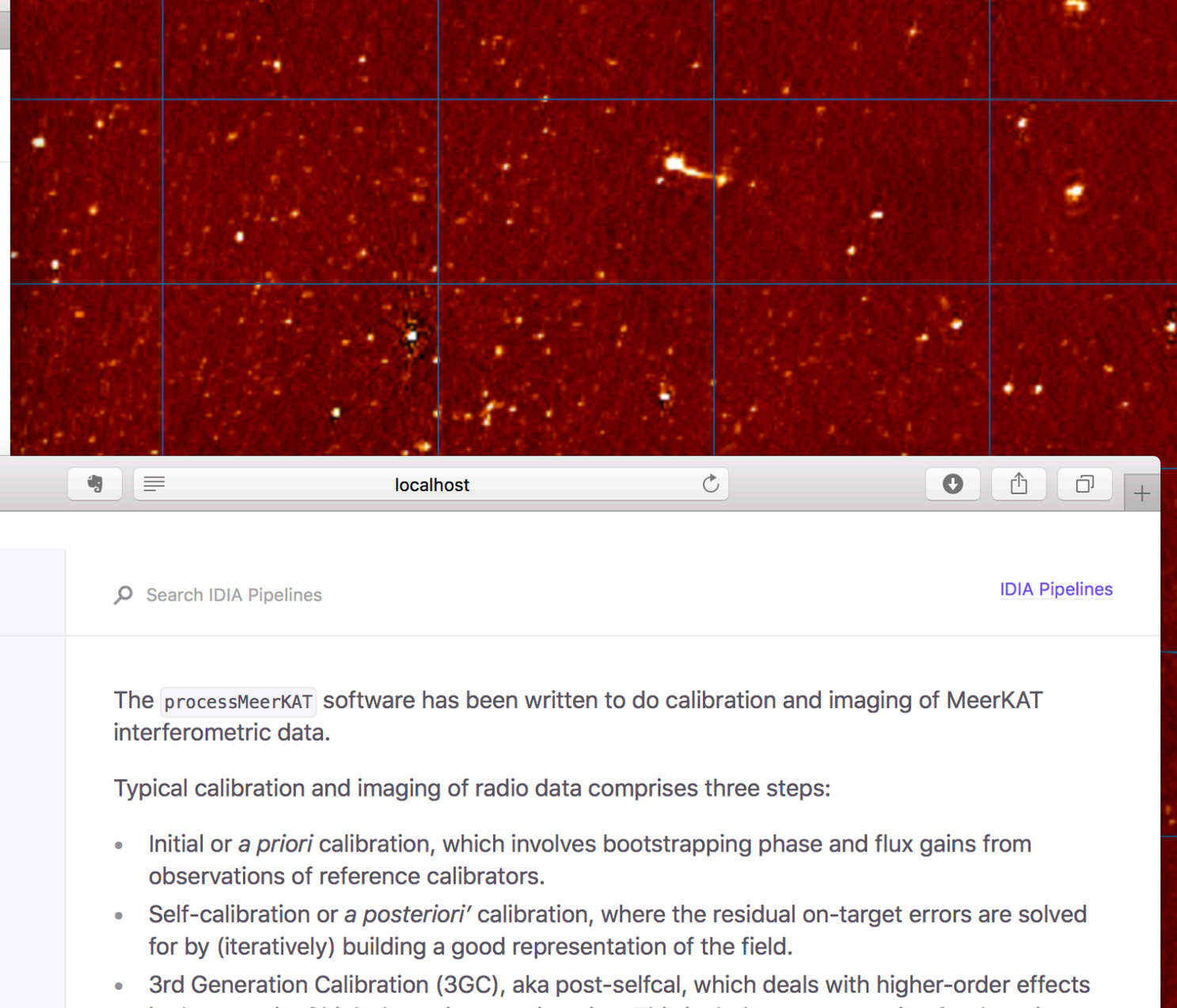
The IDIA Pipelines

The team includes memb
MeerKAT Large Survey Pr
active members who are
or the related software/al

- Jordan Collier
- Bradley Frank (Project
- Srikrishna Sekhar
- Russ Taylor

Contact

Email the project lead at
submit a ticket to our help



The `processMeerKAT` software has been written to do calibration and imaging of MeerKAT interferometric data.

Typical calibration and imaging of radio data comprises three steps:

- Initial or *a priori* calibration, which involves bootstrapping phase and flux gains from observations of reference calibrators.
- Self-calibration or *a posteriori'* calibration, where the residual on-target errors are solved for by (iteratively) building a good representation of the field.
- 3rd Generation Calibration (3GC), aka post-selfcal, which deals with higher-order effects in the pursuit of high dynamic range imaging. This includes compensating for the primary beam and direction dependent effects.

The `processMeerKAT` currently does full-polarisation *a priori* calibration on MeerKAT data, and includes automated flagging. `processMeerKAT` is written solely for the processing of data on IDIA's SLURM cluster, but future revisions will allow you to run the software on any HPC platform.

The main features of `processMeerKAT` are as follows:

- Is written in Python 3.
- Calibration algorithms only use CASA 5.4.X tasks and helper functions. ** Uses a purpose-built CASA Singularity container for parallel processing at IDIA, i.e, is fully thread-safe.
- Uses `MPICASA` to run parallel jobs over the cluster.
- Generates `SBATCH` files and ancillary helper scripts for processing.

```
localhost slots=20
10.0.0.1 slots=30
10.0.0.2 slots=40
```

```
--nologger --log2term --nogui
```

```
mpicasa -hostfile hostnames /path/to/casa --someoptions commands
```

```
mpirun
```

Executable

```
myscript.py
```

Localhost

```
orted * 20
```

10.0.0.1

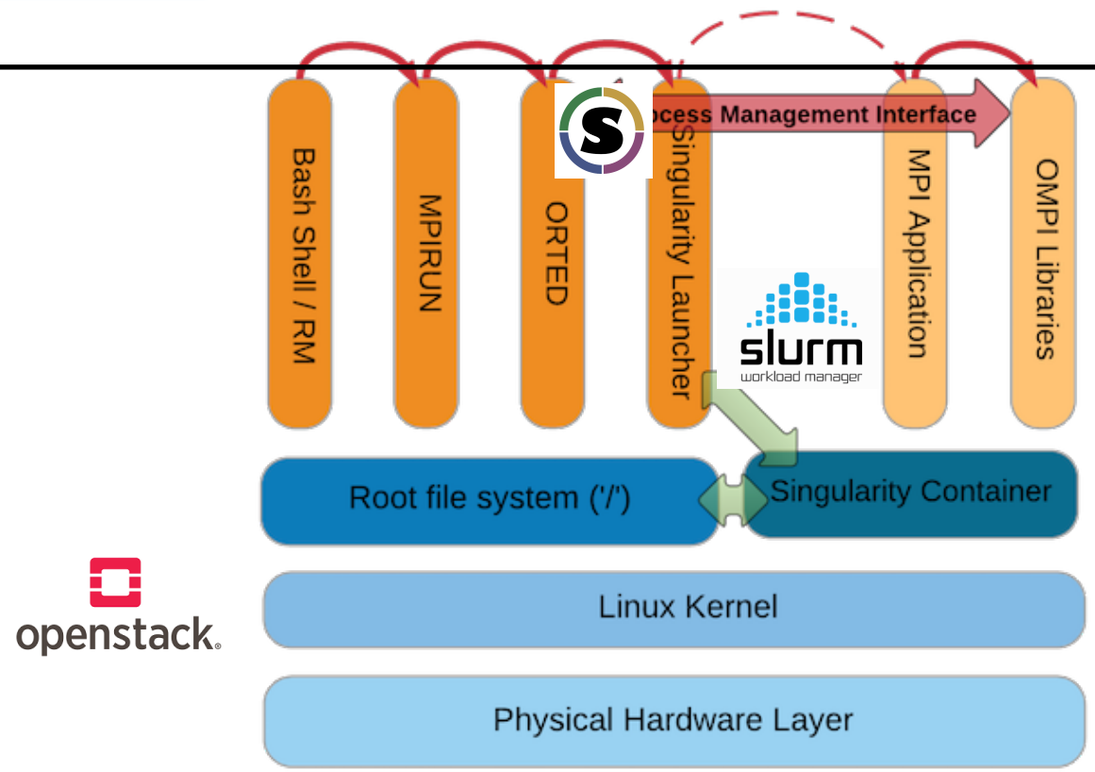
```
orted * 30
```

10.0.0.2

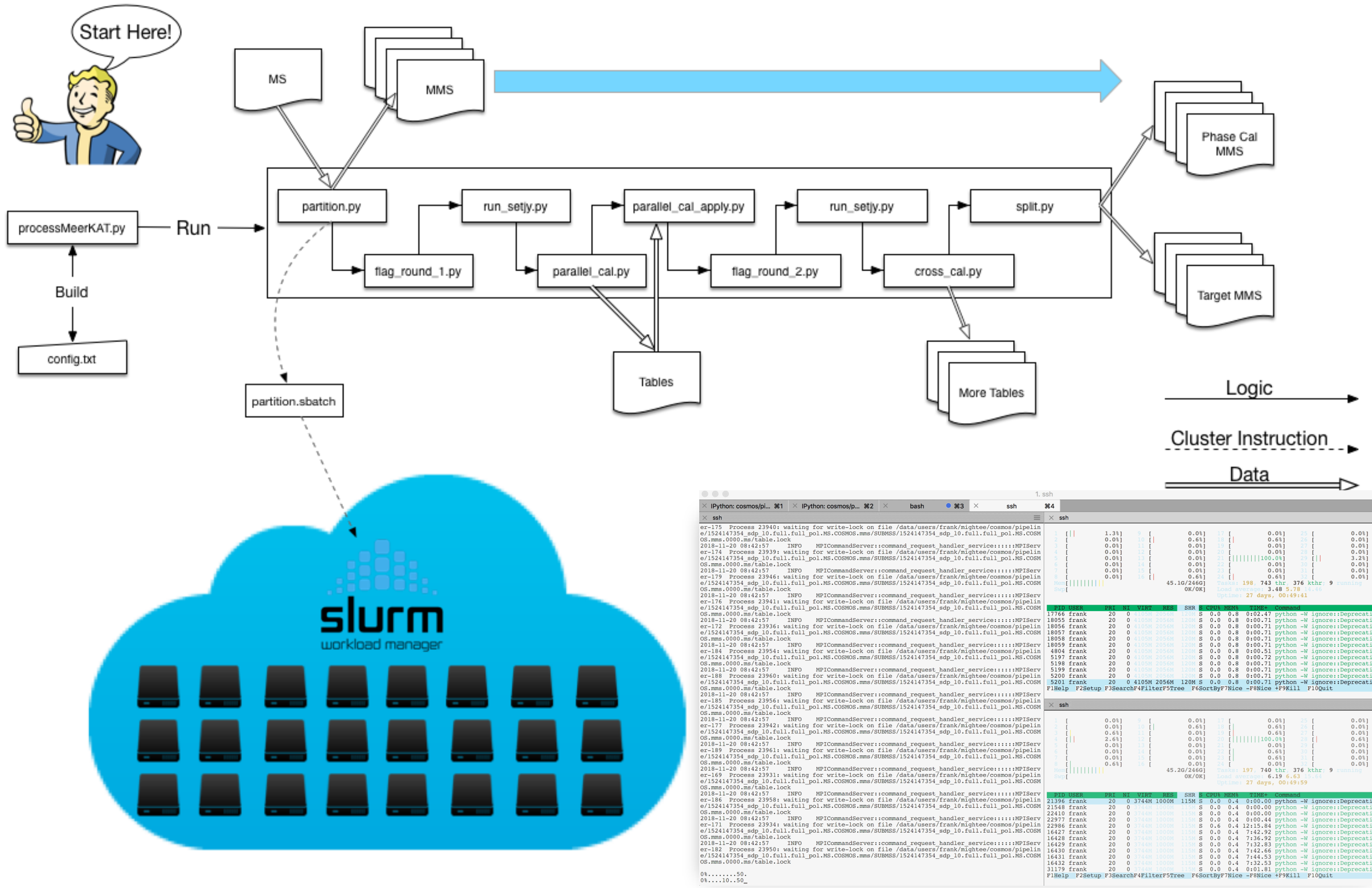
```
orted * 40
```

```
some_task(arg1='blah',
          arg2=123,
          arg4='whatever)
```

CASA



LBL

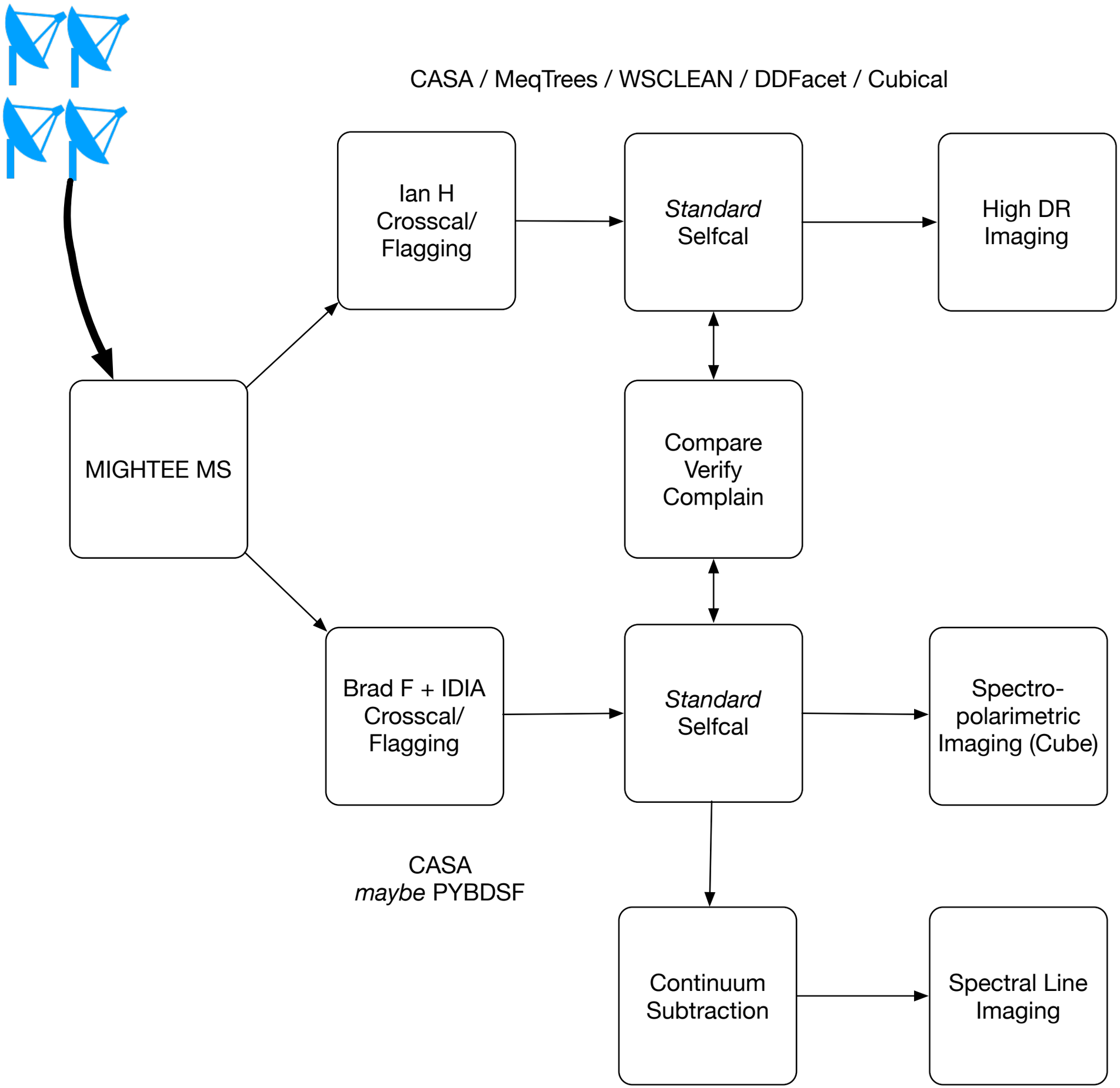

Start Here!
MS
MMS
processMeerKAT.py
Run
Build
config.txt
partition.py
run_setjy.py
parallel_cal_apply.py
run_setjy.py
split.py
flag_round_1.py
parallel_cal.py
flag_round_2.py
cross_cal.py
Phase Cal MMS
Target MMS
Tables
More Tables
partition.sbatch
Logic
Cluster Instruction
Data
slurm
workload manager
1. ssh
IPython: cosmos/pi... ⌘1
IPython: cosmos/p... ⌘2
bash ⌘3
ssh ⌘4
ssh
ssh
PID USER PRI NI VIRT RES SHR S CPU% MEM% TIME+ Command
F1Help F2Setup F3SearchF4FilterF5Tree F6SortByF7Nice -F8Nice +F9Kill F10Quit
PID USER PRI NI VIRT RES SHR S CPU% MEM% TIME+ Command
F1Help F2Setup F3SearchF4FilterF5Tree F6SortByF7Nice -F8Nice +F9Kill F10Quit
PID USER PRI NI VIRT RES SHR S CPU% MEM% TIME+ Command
26164 frank 20 0 1501M 608M 110M S 0.0 0.2 0:00.00 python -W ignore::DeprecationWarning /opt/casa-release-5.4.0
26183 frank 20 0 1501M 608M 110M S 0.0 0.2 0:00.00 python -W ignore::DeprecationWarning /opt/casa-release-5.4.0
26203 frank 20 0 1501M 608M 110M S 0.0 0.2 0:00.00 python -W ignore::DeprecationWarning /opt/casa-release-5.4.0
26137 frank 20 0 1501M 608M 110M S 44.8 0.2 6:57.06 python -W ignore::DeprecationWarning /opt/casa-release-5.4.0
27135 frank 20 0 55696 4816 4044 S 0.0 0.0 0:00.02 systemd --user
26136 frank 20 0 112M 4748 640 S 0.0 0.0 0:00.00 casa -x OMP_NUM_THREADS -x PATH -x LD_LIBRARY_PATH --prefix
26132 frank 20 0 112M 4748 640 S 0.0 0.0 0:00.07 casa -x OMP_NUM_THREADS -x PATH -x LD_LIBRARY_PATH --prefix
Active Threads
MPI/OMP Services.

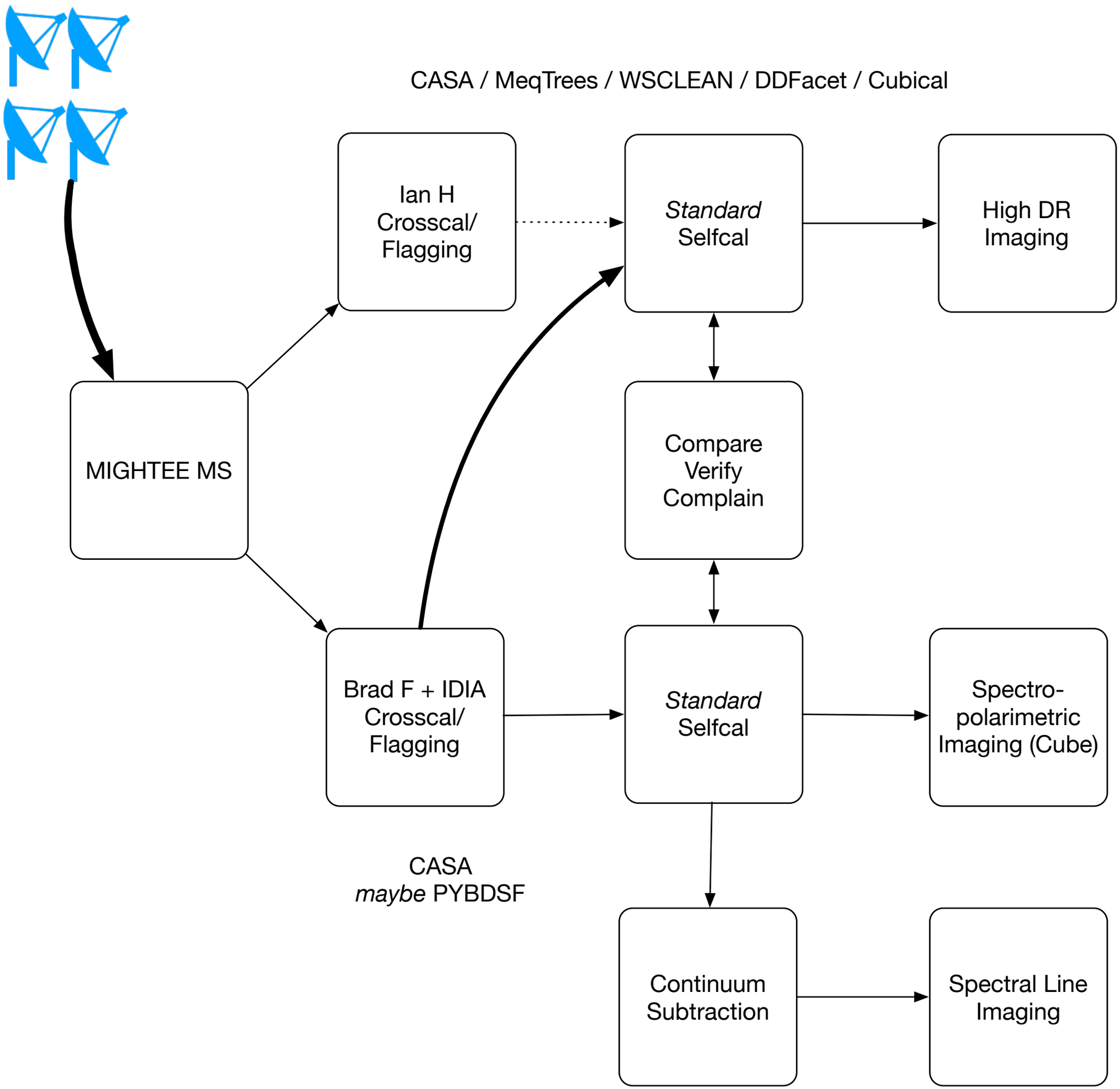
MIGHTEE

MIGHTEE Observations

DATE	ID	POINTING	TARGET	PRIMARY	SECONDARY	POL	T_int [s]	N_chan	Track [h]
2018-04-11	1523464709	COSMOS	COSMOS	J0408-6545	3C237	J1331+3030	8	4096	6.74
2018-04-12	1523518570	CDFS_16	CDFS16	J1939-6342	J0240-2309	J0521+1638	4	4096	4.26
2018-04-12	1523541036	CDFS_16	CDFS16	-	J0240-2309	J0521+1638	4	4096	4.33
2018-04-19	1524147354	COSMOS	COSMOS	J0408-6545	3C237	J1331+3030	4	4096	8.65
2018-05-06	1525613583	COSMOS	COSMOS	J0408-6545	3C237	J1331+3030	4	4096	8.39
2018-10-06	1538856059	XMMLSS_12	J0217-0449	J1939-6342	J0201-1132	J0521+1638	8	4096	8.02
2018-10-07	1538942495	XMMLSS_13	J0220-0449	J1939-6342	J0201-1132	J0521+1638	8	4096	8.07
2018-10-08	1539028868	XMMLSS_14	J0223-0449	J1939-6342	J0201-1132	J0521+1638	8	4096	8.03
2018-10-09	1539109858	ELAIS-S1_4	J0037-4359	J1939-6342	J0224-4202	J0521+1638	8	4096	8.02
2018-10-11	1539286252	XMMLSS_12	J0217-0449	J1939-6342	J0201-1132	J0521+1638	8	4096	8.05
2018-10-12	1539372679	XMMLSS_13	J0220-0449	J1939-6342	J0201-1132	J0521+1638	8	4096	8.03
2018-10-13	1539460932	XMMLSS_14	J0223-0449	J1939-6342	J0201-1132	J0521+1638	8	4096	8
2018-10-14	1539540056	ELAIS-S1_4	J0037-4359	J1939-6342	J0224-4202	J0521+1638	8	4096	8.03
									96.62

IDIA SKA AFRICA SQUARE KILOMETRE ARRAY NRF National Research Foundation


CASA / MeqTrees / WSCLEAN / DDFacet / Cubical
Ian H
Crosscal/
Flagging
Standard
Selfcal
High DR
Imaging
MIGHTEE MS
Compare
Verify
Complain
Brad F + IDIA
Crosscal/
Flagging
Standard
Selfcal
Spectro-
polarimetric
Imaging (Cube)
CASA
maybe PYBDSF
Continuum
Subtraction
Spectral Line
Imaging


CASA / MeqTrees / WSCLEAN / DDFacet / Cubical
Ian H
Crosscal/
Flagging
Standard
Selfcal
High DR
Imaging
MIGHTEE MS
Compare
Verify
Complain
Brad F + IDIA
Crosscal/
Flagging
Standard
Selfcal
Spectro-
polarimetric
Imaging (Cube)
CASA
maybe PYBDSF
Continuum
Subtraction
Spectral Line
Imaging

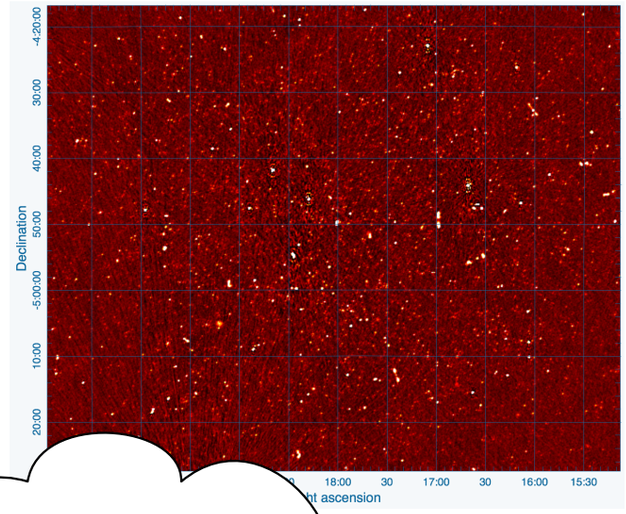
CASA / MeqTrees / WSCLEAN / DDFacet / Cubical

MIGHTEE MS

Ian H
Crosscal/
Flagging

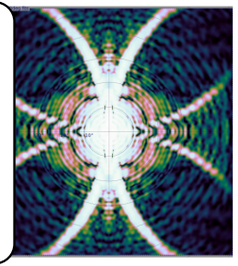
Standard
Selfcal

High DR
Imaging

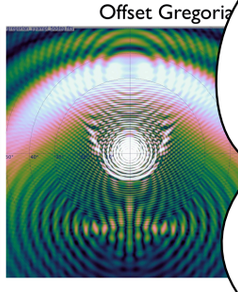


Compare
Verify
Complain

Prime Focus



Offset Gregoria



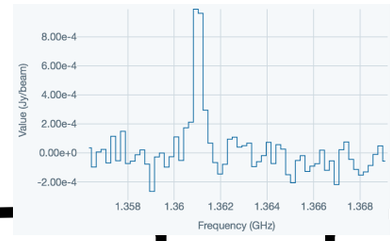
AW Imaging
MeerKAT Primary
Beam Modeling

Brad F + IDIA
Crosscal/
Flagging

Standard
Selfcal

Spectro-
polarimetric
Imaging (Cube)

CASA
maybe PYBDSF



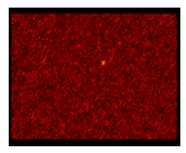
Continuum
Subtraction

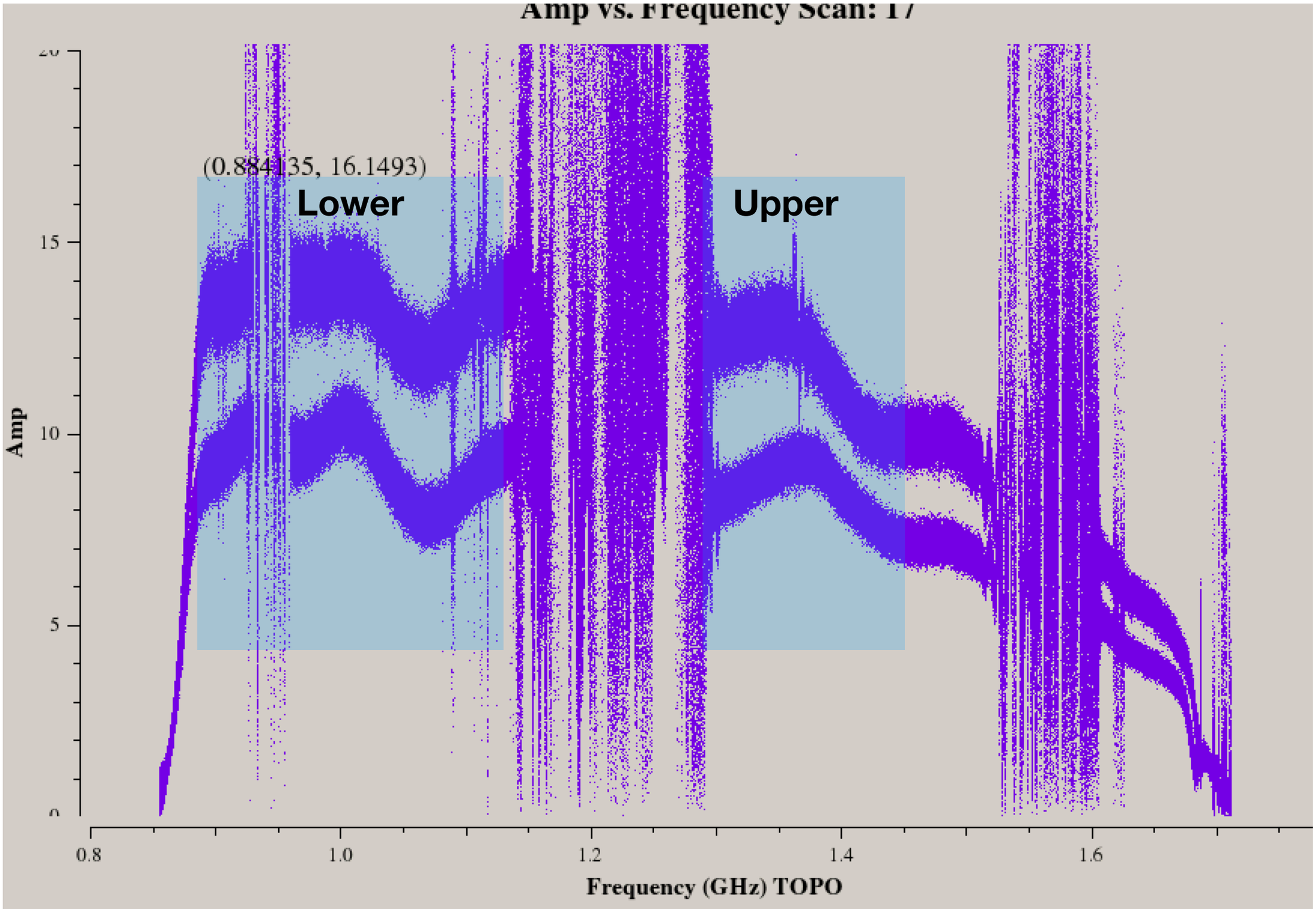
Spectral Line
Imaging

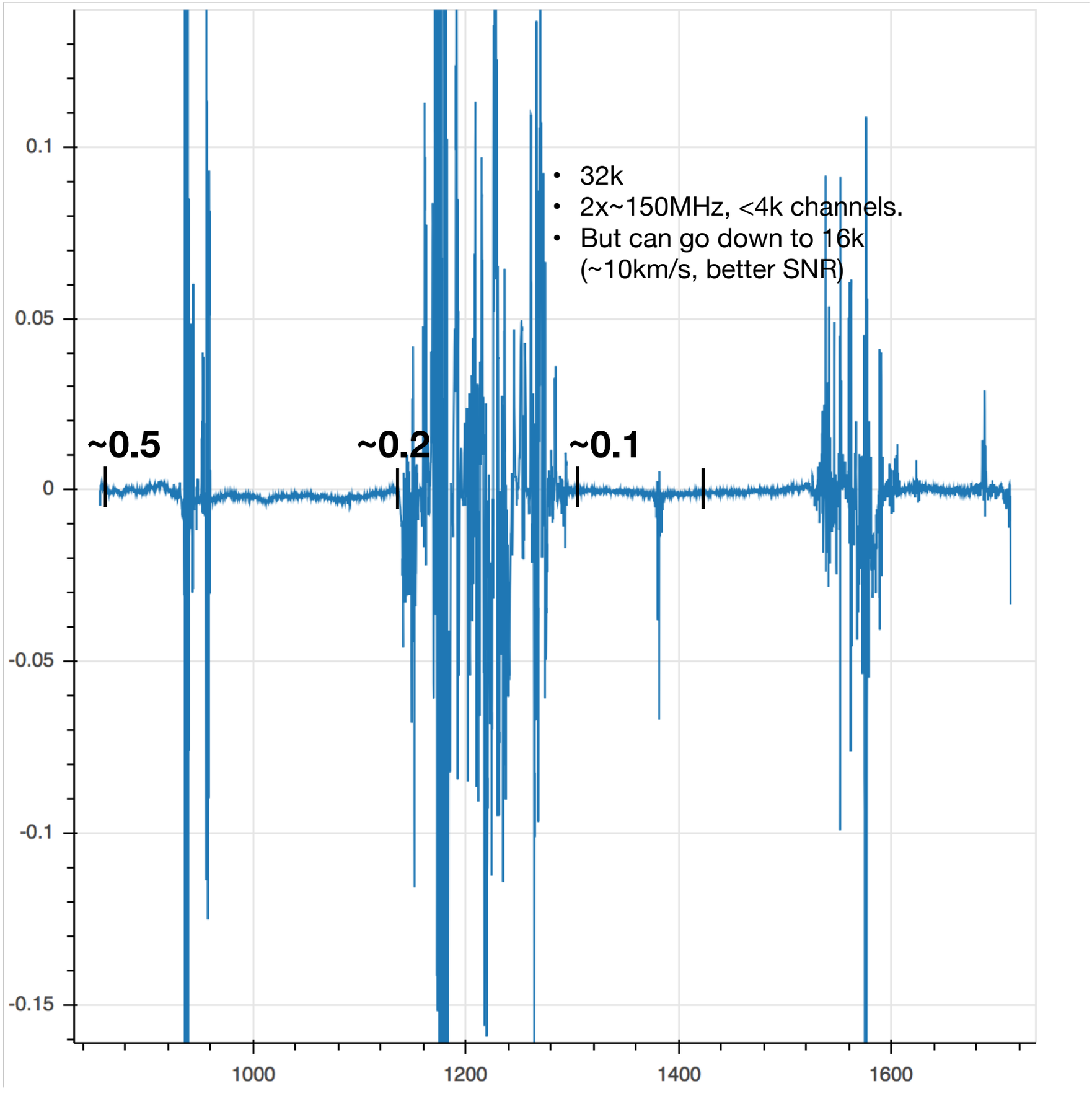
Cube(s)?

Source Finding
Validation
Science
Products

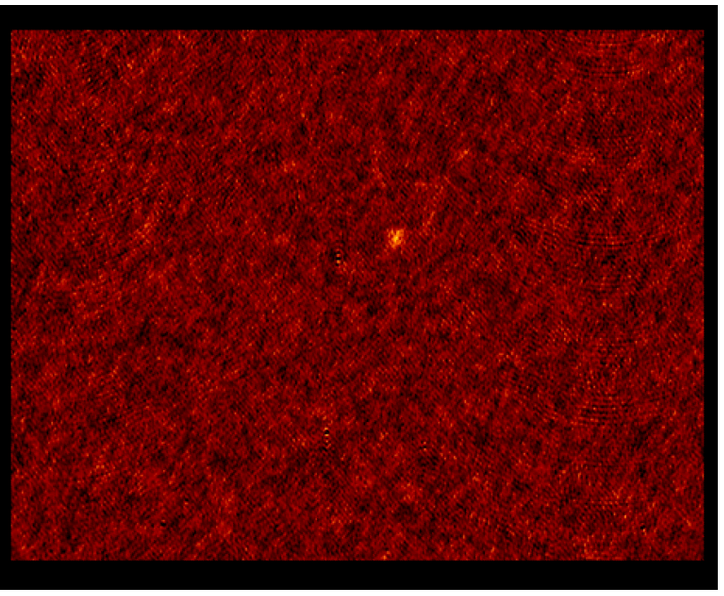
1k: ~6h

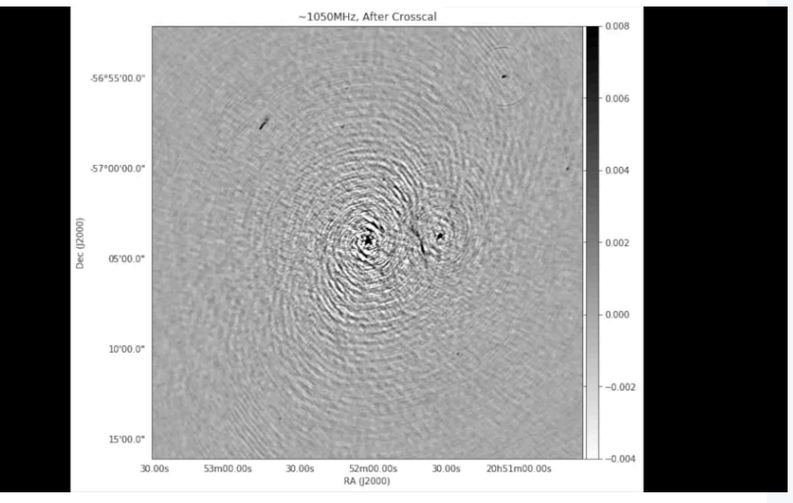


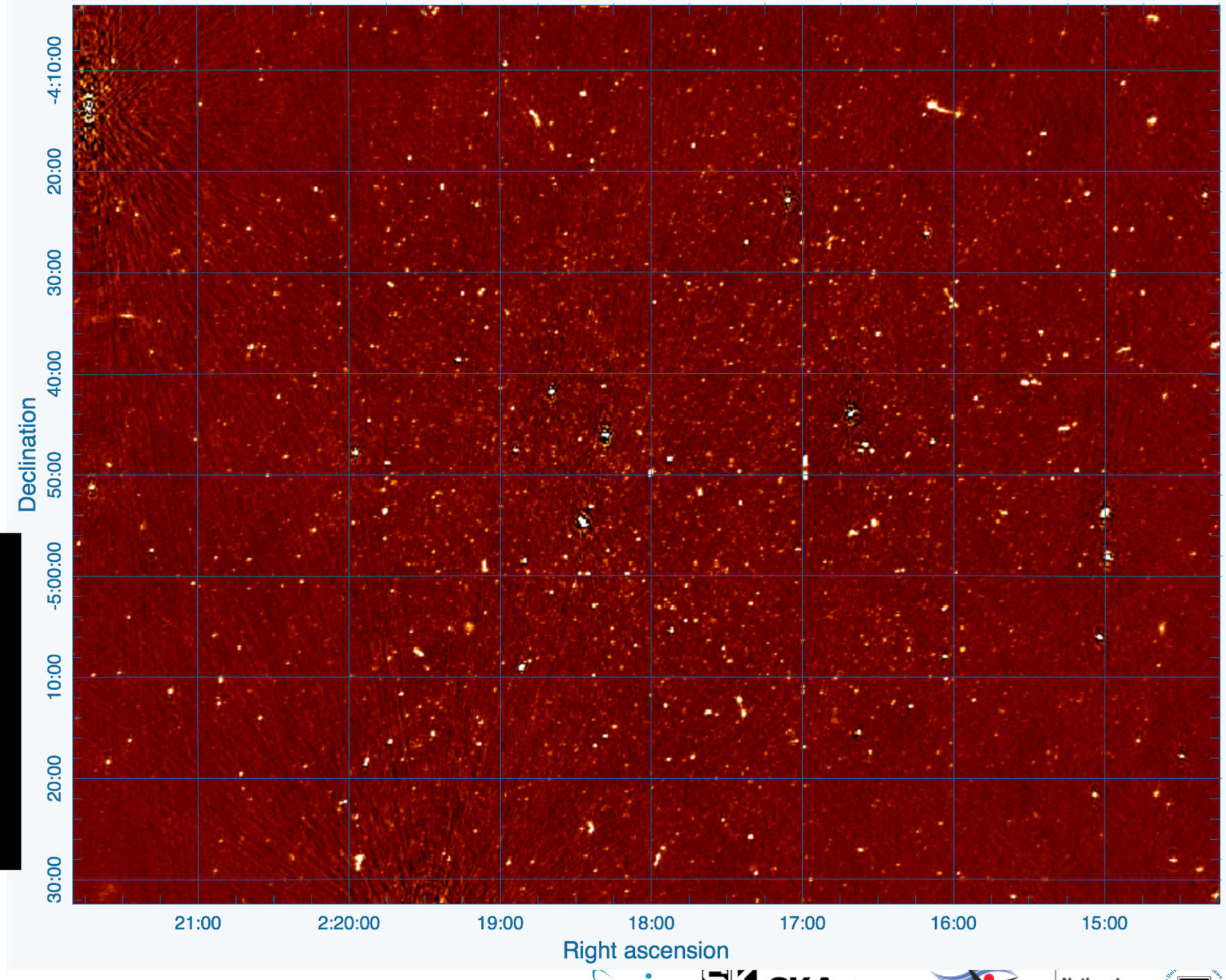

Amp vs. Frequency Scan: 17
(0.884135, 16.1493)
Lower
Upper
Amp
20
15
10
5
0
0.8
1.0
1.2
1.4
1.6
Frequency (GHz) TOPO


0.1
0.05
0
-0.05
-0.1
-0.15
1000
1200
1400
1600
~0.5
~0.2
~0.1
• 32k
• 2x~150MHz, <4k channels.
• But can go down to 16k
(~10km/s, better SNR)

Some Results

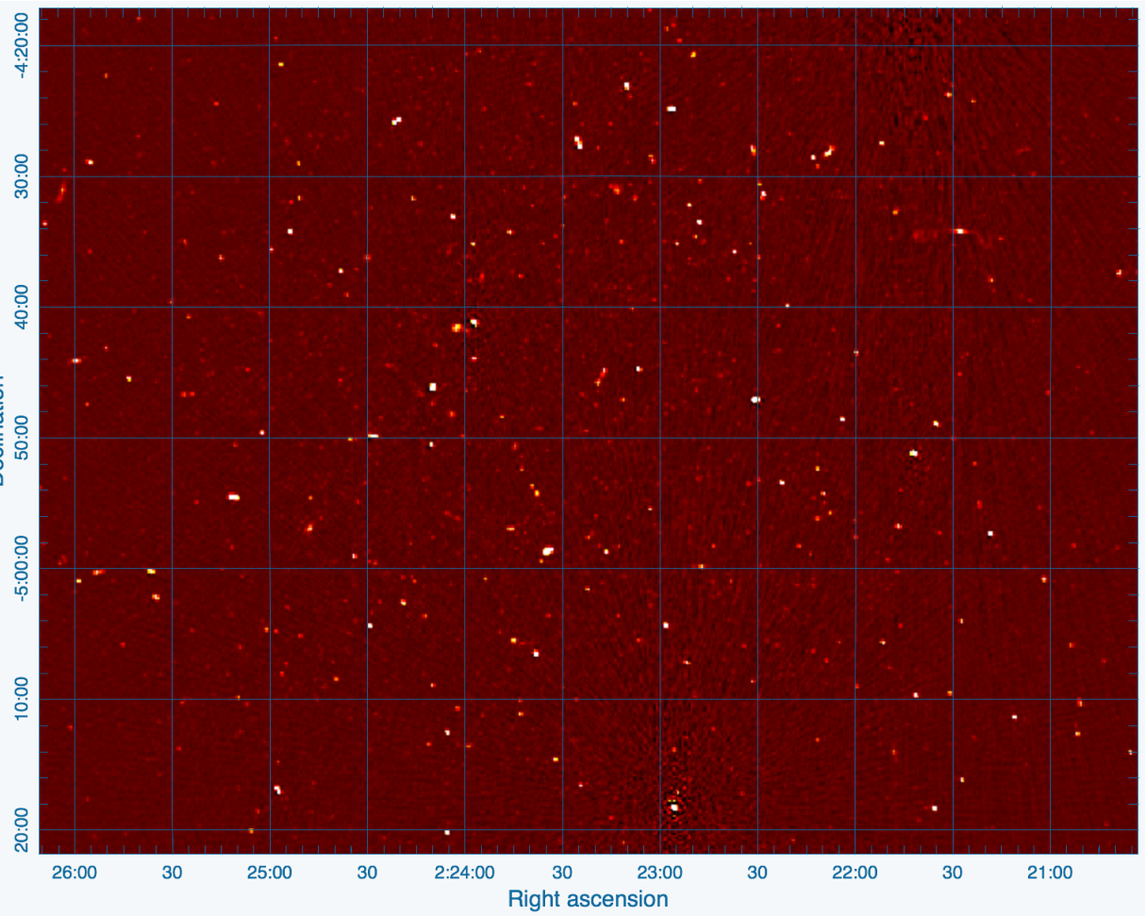


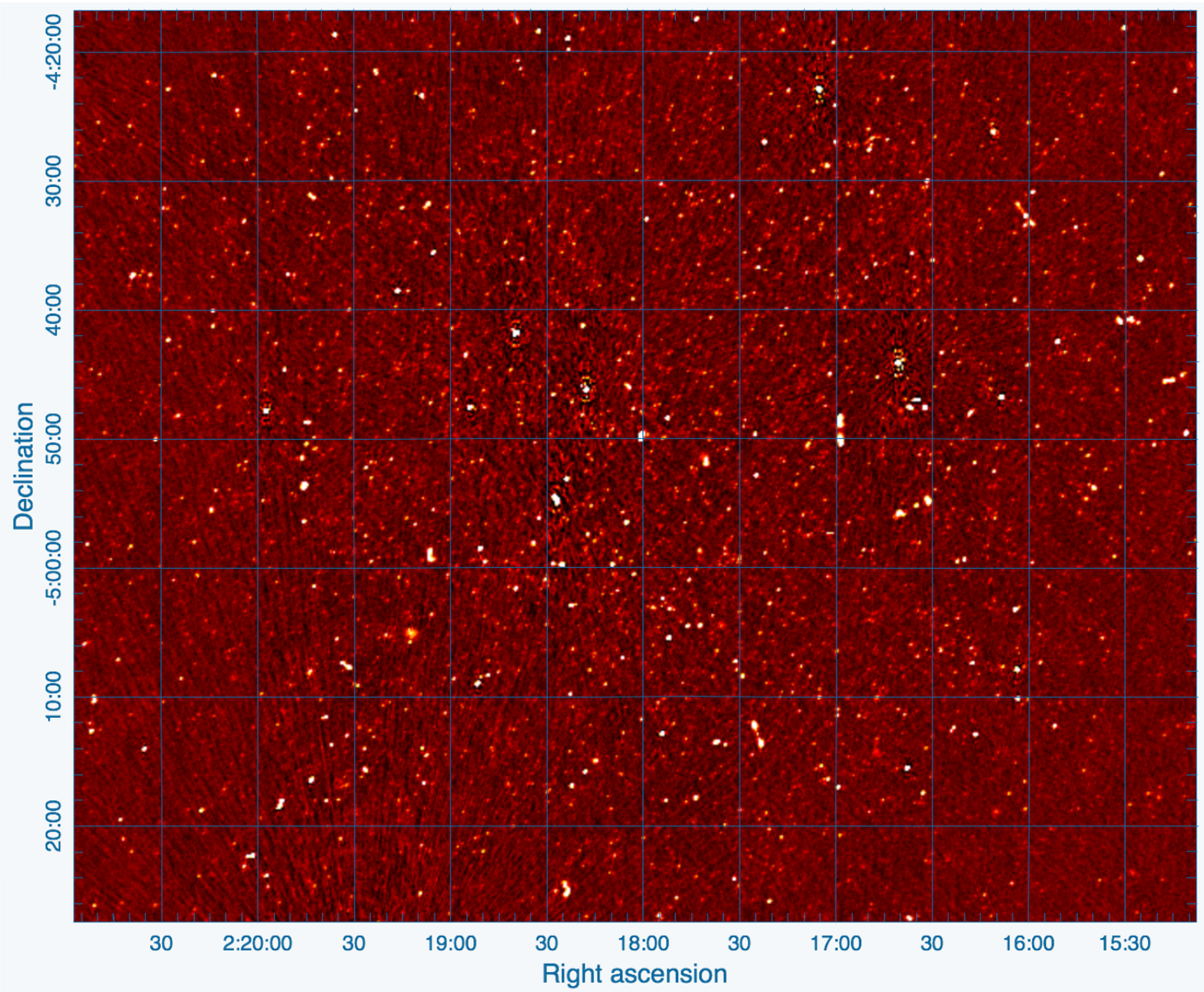




IDIA · SKA AFRICA SQUARE KILOMETRE ARRAY · NRF National Research Foundation

MIGHTEE Results





- Good imaging after Crosscal ONLY! (100MHz)

Noise ~ 20μJy/beam

- Robustness.
- Flux Scaling issues.
- Low level DDEs present.

IDIA SKA AFRICA SQUARE KILOMETRE ARRAY NRF National Research Foundation

Moving Forward

- processMeerKAT currently under performance testing.
- **Release to LSPs this month.**
- SLURM/MPICASA User Guide: released soon thereafter.
- Selfcal: April 2019 (Planned:)
- AW Projection: June/July 2019 with NRAO (Bhatnagar et al.)