Ilifu Processing and Storage for MeerKAT Projects

Brad Frank[1,2], Jordan Collier[1], Jeremy Smith[1], Russ Taylor[1]

[1]Inter-University Institute for Data Intensive Astronomy

[2] South African Radio Astronomy Observatory

February 25, 2021

Executive Summary

This document provides an overview of the Ilifu processing and storage strategy in support of MeerKAT imaging projects. We use experience running the IDIA calibration and imaging pipeline for MeerKAT 32k data sets to undertake a quantitative assessment of requirements and data policies for successful processing of MeerKAT observations from raw data to science data products. This strategy informs the data management plans and resource estimates for MeerKAT projects. The document is intended as background information for MeerKAT project members developing data processing plans on Ilifu.

Contents

1 Introduction

The MeerKAT telescope will conduct several projects in anticipation of the SKA. The majority of these focus on interferometric imaging, using both the continuum and spectral-line capability of MeerKAT over the three nominal bands available. While there are many scientific commonalities between the imaging projects, they each have unique technical requirements which in turn present knock-on requirements for processing. That is, there is a strong link between the realisation of scientific results for these projects and the associated technical operations of pipelines and processing facilities.

In preparation for MeerKAT Large Survey Projects, many imaging projects have developed internal designs for calibration and imaging workflows, and some have been able to commission these plans using early science data from MeerKAT. The availability of early science data and Open Time projects have also provided the opportunity to assess the performance of calibration and imaging workflows. As a consequence, the MeerKAT community is presented with an opportunity to compare their initial processing plans (which are based on science goals) with the real-world performance of workflows on currently operational processing facilities. With a slightly more detailed and rigorous approach, such a comparison could highlight scientific and technical

plans that need revision, and provide a framework to plan a way forward that best matches the requirements of the MeerKAT projects.

This document uses the experiences of the IDIA (Inter-University Institute for Data Intensive Astronomy) Pipelines Team as a starting point to initiate a dialogue between the MeerKAT project members and IDIA, with the ultimate aim of aligning the technical requirements of imaging projects and the functional performance and operation of the IDIA pipeline, and other similar workflows, on the South African Ilifu data intensive cloud facility and the associated services.

The IDIA pipeline takes advantage of the distributed computation available on Ilifu through concurrent and parallel processing. However, we expect the projections and discussion points presented in this document to be both instructive and generally applicable irrespective of pipeline implementation. The basic transfer, processing and storage requirements are generic, since many calibration and imaging pipelines use of similar underlying software packages for cross-calibration and (in some cases), self-calibration.

This document is structured as follows. We provide an overview of the kind of processing (either partially or completely) that will take place at Ilifu; we provide an overview of the Ilifu facility, the breakdown of data products and a description of the IDIA Pipeline. We then present processing and operational models and recommended approaches and data policies for processing of MeerKAT observations on Ilifu.

2 MeerKAT Processing IDIA/Ilifu

2.1 MeerKAT Processing

Most projects using ilifu will perform full-scale science processing on the cluster, utilising Ilifu for all stages of processing. Many projects plan to use the IDIA processMeerKAT pipeline[1] (see Section 4.1), or the CARACal[2] pipeline, or the Oxkat[3] pipeline. Many projects will also make use of direction dependent calibration tools[4].

Some MeerKAT projects will use a combination of Ilifu and their own resources, with Ilifu as an initial staging ground for transfers/pre-processing, including data selection and averaging, quick-

[1] https://idia-pipelines.github.io/docs/processMeerKAT
[2] https://github.com/caracal-pipeline/caracal
[3] https://github.com/IanHeywood/oxkat
[4] see https://github.com/saopicc

look images and basic quality assessment, and transferring of data to other facilities.

2.2 Ilifu Cloud System

A primary goal of IDIA is to build capacity and expertise in data intensive research at South African universities to support MeerKAT science projects, specifically servicing five of the eight MeerKAT Large Survey Projects, and a number of MeerKAT Open Time Projects and Director's Discretionary Time projects. IDIA manages the Ilifu research cloud infrastructure which supports both astronomy and bioinformatics South African research communities. The Ilifu facility provides both compute and storage resources and offers infrastructure, platform, software and support as a service. Data transfer nodes and services provide fast data transfers from the SARAO archive. A new client-server astronomy visualisation tool, the Cube Analysis and Rendering Tool for Astronomy (CARTA)[5], allows for web-based, efficient interactive visual analytics of large astronomy data sets hosted on Ilifu.

The Ilifu cluster is the primary infrastructure for data processing and includes a pool of compute resources and access to storage. Using the SLURM job scheduling software, the cluster supports batch job submission, for running data processing pipelines, and workflows. A development and analytics environment is implemented through a Jupyter notebook and JupyterLab service (jupyter.ilifu.ac.za) with resources selected by the user to fit the scale of the task. Several Jupyter kernels are available with common astronomy software environments for processing and analytics, e.g. CASA, source finding, machine learning, plotting, etc.

At the time of writing this report, the Ilifu infrastructure includes the following compute and storage resources:

Compute

- 110 x compute nodes, 32 CPUs, 256 GB RAM
- 2 x compute nodes, 32 CPUs, 512 GB RAM
- 4 x GPU nodes, 32 CPUs, 256 GB RAM, 2 x Tesla P100 16 GB GPU

Storage

- 400 TiB BeeGFS (scratch storage)
- 2.9 PiB CephFS

[5] https://cartavis.github.io/

Of the 110 compute nodes, 88 nodes are included in the Ilifu SLURM cluster, with 76 nodes in the main partition and 12 nodes in the Jupyter partition. The number of nodes within the cluster and allocated to partitions is subject to change as the nodes are moved between partitions and other services as needs demand. Additional CephFS storage of approximately 2.8 PiB will be available by October 2020. The CephFS storage values indicate usable storage and are calculated as 70% of raw available storage to allow for redundancy and file-system overheads. Usable storage is intended to operate at below 80% capacity for the storage to operate efficiently.

In addition to the above infrastructure, Ilifu also provides a well developed software ecosystem. Software environments are encapsulated using Singularity[6] containers. The containers allow for software stacks to be made available across distributed infrastructure. This allows for easy access to and sharing of custom software stacks, provides a flexible software environment and supports reproducible science.

3 Data

3.1 Raw data properties and volumes

The visibility data from MeerKAT is stored at Ilifu in CASA MeasurementSet (MS) format. The data volume is a function of the observation length, bandwidth, number of polarisations, number of channels and time dump interval. A typical L-band data set without data filtering or averaging will be an 8 hour observation with 32,768 frequency channels over a 856 MHz bandwidth, with 64 antennas, all four polarisations (HH,VV,VH,HV), and an 8 second dump / integration time. Data from the UHF-band and S-band will have the same data volume if the number of channels and dump times are the same. The MS includes four large columns (data, flag, flag category, and weight spectrum), and a number of other smaller columns.

The data volume of a column in the MS can be calculated in TiB using Equation 1.

$$V_{col} = t_{obs} \times N_{baseline} \times (\frac{3600}{d}) \times \frac{b}{8 \times 1024^4} \tag{1}$$

where

$$N_{baseline} = \frac{N_{ant} \times (N_{ant} - 1)}{2},$$

is the number of baselines given by the number of antennas N_{ant} (excluding auto-correlations),

[6] https://sylabs.io/

t_{obs} is the observation length in hours, d is the dump time in seconds, and b is the number of bits per row of the column.

For the four large columns of a raw MS, b can be calculated for each of the columns as follows: 1) data columns, $b_{data} = N_{chan} \times N_{pol} \times 64$; 2) flag column, $b_{flag} = N_{chan} \times N_{pol} \times 8$; 3) flag category, $b_{category} = N_{chan} \times 8$; and 4) weight spectrum, $b_w = N_{chan} \times N_{pol} \times 32$. Using $t_{obs} = 8h$, $N_{ant} = 64$, $N_{pol} = 4$ and $d = 8s$, the volume of a typical raw MS (including a single data column, flag and weights columns) is

$$V_{raw} = V_{data} + V_{flag} + V_{weights} = 7.14 + 1.12 + 3.57 \text{ TiB} = 11.83 \text{ TiB} \tag{2}$$

This is the size of a raw MS data set and does not include corrected or model data columns that are produced during the calibration process. For the simplicity of this analysis we use this value as the reference volume, $V_{ref} = 12$ TiB. For a given observing and data selection the data volumes can be estimated using Equation 3.

$$V \approx 12 \left(\frac{t_{obs}}{8} \times \frac{\Delta\nu}{856} \times \frac{P}{4} \times \frac{1}{N} \times \frac{8}{t} \right) \quad \text{TiB} \tag{3}$$

where t_{obs} is the observation length in hours, $\Delta\nu$ is the bandwidth in MHz, P is the number of polarisations, N is the number of frequency channels averaged, and t is the bin width for time averaging in seconds.

The volumes of the imaging data products are given by $b = 32$ bits per pixel, calculated in GiB as:

$$V = N_{pix} \times N_{pol} \times N_{chan} \times \frac{32}{8 \times 1024^3} \text{ GiB}, \tag{4}$$

where N_{pix} is the image size, N_{pol} is the number of Stokes channels (> 1 for a Stokes cube), and N_{chan} is the number of frequency channels (> 1 for a spectral-line or continuum cube). For a typical Multi-Frequency Synthesis (MFS) image size of 4096×4096, and cube image size of 2400×2400, the respective volumes of a single MFS image, a 350 channel Stokes cube, and a 21,000 channel spectral-line cube (without channels above 1420 MHz or below 880 MHz) are

given by:

$$
\begin{aligned}
V_{\mathrm{Spectral}} &= 2400^2 \times 1 \times 21,000 \times \frac{32}{8 \times 1024^3} = 451\,\mathrm{GiB} \\
V_{\mathrm{Stokes}} &= 2400^2 \times 4 \times 350 \times \frac{32}{8 \times 1024^3} = 30\,\mathrm{GiB} \\
V_{\mathrm{MFS}} &= 4096^2 \times 1 \times 1 \times \frac{32}{8 \times 1024^3} = 0.1\,\mathrm{GiB}
\end{aligned}
\tag{5}
$$

3.2 Usable L-band data

Experience has shown that about 20-30% of the MeerKAT L-band RF is continuously occupied by strong, persistent RFI. Figure 1 shows a plot of the mean visibility amplitude for a typical observation of the primary flux calibrator J1939-6342 from the MIGHTEE project. The regions occupied by strong RFI are obvious. These regions are consistently occupied by RFI for all observations to date. The two ranges from about 1170–1300 MHz and 1530–1630 MHz are dominated by satellite navigation systems. The small band from about 930 to 960 MHz is allocated to mobile systems and aeronautical navigation.

The four segments of the spectrum that lie outside these regions are relatively quiet. These four bands are labelled bands A, B, C and D, in the Figure and represent about 70% of the observable RF. The frequency limits of the bands are listed in Table 1. Narrow-band low-level RFI can be seen within these bands in the difference spectra (blue). This level of RFI can be largely removed by automated flagging.

Band	Frequency Range
Band A	880 – 933 MHz
Band B	960 – 1163 MHz
Band C	1299 – 1524 MHz
Band D	1630 – 1680 MHz

Table 1: Useful segments of the MeerKAT L-band RF for spectral processing. See Figure1.

The strength of the RFI signal depends on baseline lengths. Experience has shown that there is retrievable data in the RFI affected regions on baselines longer than about 600m. For the purpose of broadband continuum multi-frequency synthesis imaging, there is useful data on these longer baselines. Therefore the full RF band is processed for the continuum imaging. However, spectral channels outside of bands A, B, C and D do not contain sufficient data for

quality channel images, so only data within these four bands are retained by default for spectral line and spectro-polarimetric processing.

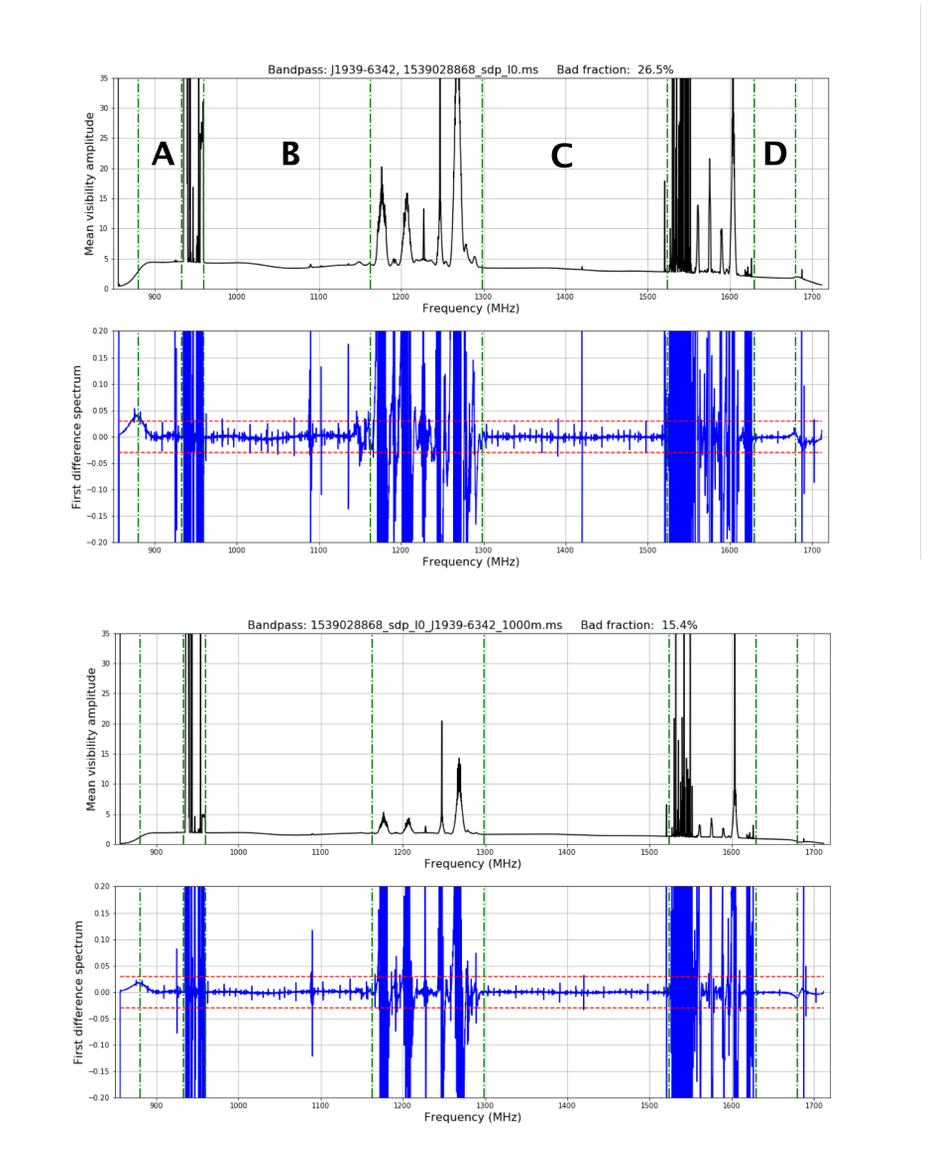

Figure 1: RFI occupancy of the MeerKAT L-band RF. The panels show the mean visibility amplitude versus frequency for an observation of J1939-6342 in black and the first difference (derivative) spectrum in blue. The top panel shows data for all baselines. The bottom panel for baselines longer than 1000 meters. Four sections of the band, representing about 70% of the data, lie outside of regions of strong persistent RFI. These four bands are delineated by the vertical green dashed lines and labelled A, B, C and D.

4 Processing Overview

MeerKAT astronomers and IDIA staff have experience running several radio astronomy processing tools on the Ilifu system. In addition to running CASA-based scripts, the following pipelines and packages have been successfully used on the system:

1. IDIA Pipeline (details provided below)
2. CARACal[7]
3. OxKAT[7]

Each pipeline or package has the associated software container or environment available on Ilifu. In this section, we focus on the IDIA Pipeline for various use-cases as an illustrative example.

4.1 The IDIA Pipeline

The IDIA pipeline is designed to provide a flexible, modular calibration and imaging pipeline to process MeerKAT observations from raw visibilities to image data sets. It is based on the `CASA` software package, and is designed to exploit parallel processing with `MPI` and concurrent processing via the SLURM job management system. Pipeline processes are deployed in arrays of distributed virtual machines running purpose-built `CASA` singularity containers.

Each stage of the pipeline is controlled by a python script. Default scripts are executed from the central pipeline repository. Users of the pipeline can easily customise the pipeline modules by providing their own local versions of scripts, or edited versions of the pipeline scripts. Execution of the jobs within the distributed environment is managed by the pipeline architecture, transparent to the user.

The pipeline is designed to serve three science use cases.

1. Extragalactic HI spectral line science by the creation of continuum-subtracted high spectral resolution total intensity cubes.
2. Polarisation science by the creation of low spectral resolution full-stokes, spectro-polarimetry cubes.

[7] Users intending to use these packages should contact the respective developers and IDIA support staff for notes on implementation.

3. High sensitivity continuum science by the creation of broadband multi-frequency synthesis total intensity images. For the MIGHTEE project calibrated data are transported to Oxford or Rhodes University for direction-dependent corrections and MFS imaging.

The high level schematic of the intended pipeline modules and data flow, for the example use case of the MIGHTEE survey, is shown in Figure 2. Raw data is transported to the Ilifu cloud over a dedicated 10 Gb/s fibre from the SARAO archive at the Centre for High Performance Computing (CHPC). The data are partitioned into several sub-bands for concurrent processing through the first stage of the pipeline. Following RFI excision, total intensity only full-spectral resolution data sets and full-polarisation low spectral resolution data sets are created. Following partition, standard full-polarisation calibration is carried out on the low frequency resolution data. This is done in two rounds, with an additional RFI excision step after the first round. Gain tables for bandpass, polarisation leakage, time-dependent gains and H-V phase are derived from observations of calibration sources. The calibration tables are applied to both low and high spectral resolution data. A high resolution bandpass calibration table is then derived for the high resolution data set to remove residual high resolution bandpass effects.

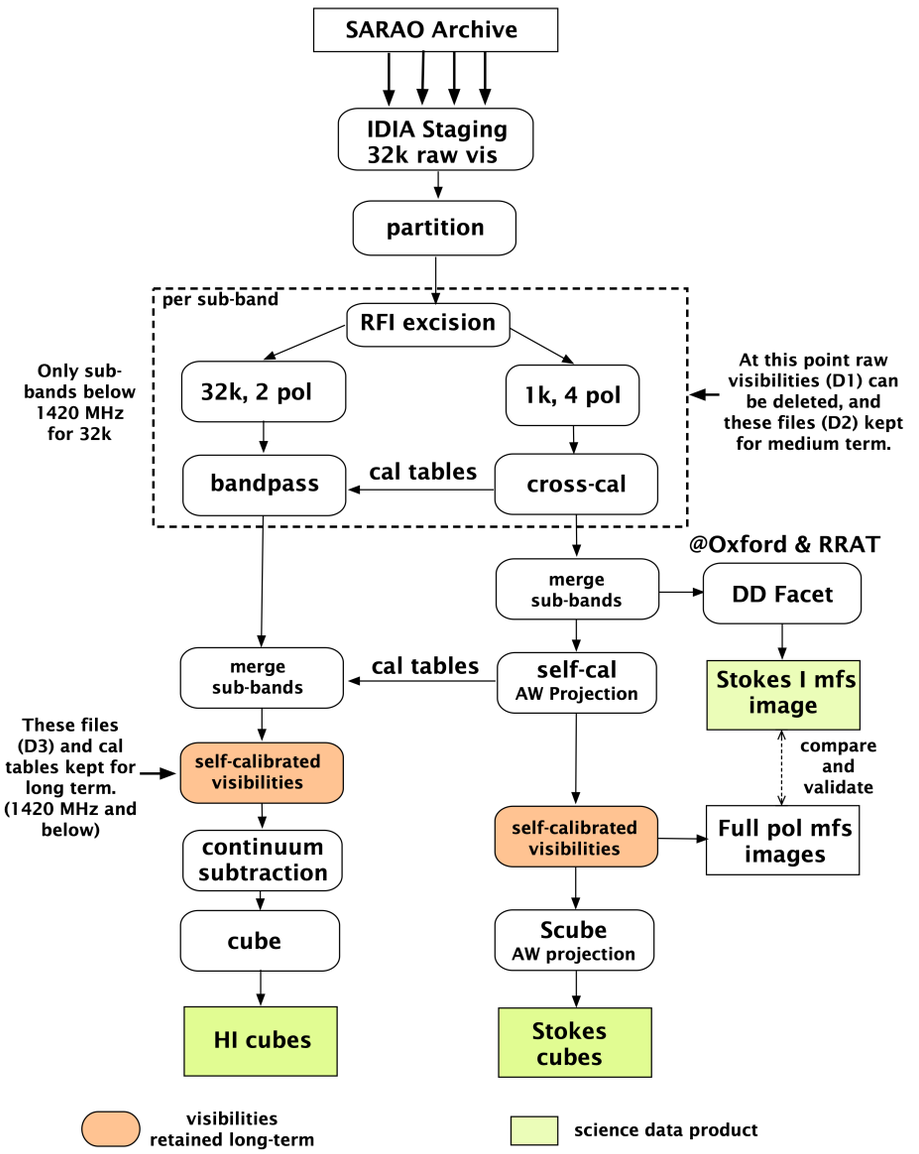

Figure 2: Schematic of the intended IDIA pipeline process flow, for the example use case of the MIGHTEE survey.

Following the cross-cal stage, the calibrated sub-bands are merged to full band data sets. The low-resolution data set is run through total intensity direction dependent self-calibration with

DDFacet[8] to create a broadband MFS image by a teams at Oxford and Rhodes University.

For the polarisation science product, the low frequency resolution data are self-calibrated with AW-projection to remove *a priori* direction-dependent beam and leakage errors from the primary beam direction-dependent Mueller matrix (leakage beams). Spectro-polarimetric cubes are generated using frequency-dependent AW-projection from the self-calibrated visibilities.

For HI science, the calibration tables are applied to the high spectral resolution data, and continuum-subtracted spectral line cubes are generated at full spectral resolution. Since extragalactic HI science does not make use of frequencies above the HI rest frequency, the HI use case retains only the data below 1420 MHz (bands A, B and about 60% of C).

4.2 The processing time equation

To parameterise the processing time through the pipeline, we break down the workflow into the following stages and assign a coefficient for each:

- Data transfer (A_{TR})
- Partition (A_{PA})
- Cross-calibration (A_{CC})
- Self-calibration (A_{SC})
- Science Imaging (A_{SI})

where the coefficients are defined as the processing time for each stage divided by the observing time T_{obs}, such that the total processing time is given by

$$T_{proc} = (A_{TR} + A_{PA} + A_{CC} + A_{SC} + A_{SI}) \cdot T_{obs} \tag{6}$$

As noted in section 3.1, an observation has a typical duration of about 8 hours.

Equation 6 assumes that the each stage of the pipeline, and hence the total processing time T_{proc}, scales linearly with T_{obs}. This is the *ideal* case and any estimates based on equation 6 represents the minimum time expected to process a specific volume of data, since the finite wait/queue time for compute resources between each step is not included in this equation.

[8] https://github.com/saopicc/DDFacet

We estimate typical values of the coefficients for the processing steps listed in Equation 6 for a typical observation using the 32k correlator mode over $T_{obs} = 8\,\mathrm{hours}$, with all the available correlation products (XX, XY, YX and YY), all 64 antennas, and an 8 second dump time. For smaller data sets (e.g. observed in 4k correlator mode), the coefficients will reduce accordingly.

The coefficient values for these observing parameters are estimated on the basis of current experience with the data transfer throughput, benchmarking of the IDIA Pipeline over an 8 node allocation, and our experience with science imaging on the Ilifu cluster, and are given by: $A_{TR} = 3.0$, $A_{PA} = 1.0$, $A_{CC} = 0.05$, $A_{SC} = 1.0$, $A_{SI} = 3.0$. This presumes that the cross- and self-calibration steps are done on the continuum data. Science Imaging produces both spectral line, and continuum/polarisation data products. However, post-processing analysis of calibrated data is not included in the estimate (e.g. mosaicking and image analysis). Once the data are at Ilifu, the end-to-end T_{proc} resulting from the use of these coefficients is approximately 40 hours.

This makes the simplifying assumption that the spectral-line, and continuum/polarisation branches of the pipeline are done simultaneously over a nominal resource request, and that the run time of that particular module is limited by the longest (or the most computationally expensive) part of the pipeline.

For example, we assume that spectral line, and continuum/polarisation imaging are done simultaneously, but that the time-to-completion for "Science Imaging" is constrained by whichever process takes the longest. We also assume that each module requests the same amount of resources.

4.3 Data Management

During the course of processing an observation, several versions of visibility data are created at various stages. The data management plan attempts to minimise the amount and length of time that large and intermediate data sets are stored on disk, while at the same time allowing for flexibility for experimentation in processing. Data that is necessary for downstream processing is retained for the long term.

The data flow through the pipeline for a single observation produces the following visibility data stages:

D1: Raw visibility MeasurementSets transferred from the SARAO archive.

D2.1: Pre-processed visibility data that has initial (conservative) flagging for RFI, before averaging. Two versions are created: a full-spectral resolution version containing only the two

parallel hand correlations, and a low-spectral resolution version with all four correlations.

D2.2: Visibility data partitioned into Multi-MeasurementSets (MMSs) in N sub-bands for concurrent processing through the cross-cal stage. After cross-cal, D2.2 includes the corrected and model data columns, and is inflated by $\sim$2.5x compared to D2.1.

D3: Calibrated visibility data with the sub-bands merged into a single MMS. This data set is significantly smaller than D2.2 as it contains only a single "data" column.

D4: Calibration tables and image data products.

For spectral-line data, D2.1 should remove everything above $\sim$1420 MHz, and perhaps some of the lower frequency band (e.g. when observing nearby galaxies). Projects with no polarisation component should remove HV and VH correlations for D2.1. All of this can be done by configuring data transfers from the SARAO archive to select or average data[9], which pre-processes the data such that D1 effectively becomes D2.1. Once data set D2.1 is created, D1 should be removed. Apart from special use cases, D2.2 should be removed almost immediately, with only the calibration tables and flag versions stored long-term. Data that needs reprocessing can begin again with D2.1, and possibly apply the calibration tables and flags. Once the cross-cal and self-cal process has been validated, D2.1 should be removed as soon as possible. In practice, D2.1 can be retained for a short period to commission and develop calibration and imaging strategies – retaining D2.1 avoids the re-transfer of D1 if potential bugs or enhancements are found during a notional period of quality assurance (QA) during the initial phases of observations. D3 and D4 will be retained long-term for downstream processing, such as combining visibilities for multiple observations of the same pointing, visibility-plane mosaicking of multiple pointings, or for scientific analysis using the visibility data. D3 can be retained as long as a project is active and using the data. Ilifu does not however provide long term archiving of data. A separate plan for archiving of D3 data products should be made once a project is completed. For South African project data, Ilifu plans to coordinate a data repository system as part of the national South African DIRISA cyber-infrastructure strategy.

The intended image outputs of the pipeline are:

1. A broadband multi-frequency synthesis, high-sensitivity total intensity continuum image
2. Full-spectral resolution continuum-subtracted, total intensity 3D data cubes
3. Low-spectral resolution, full-Stokes 4D image cubes.

[9] See http://docs.ilifu.ac.za//astronomy/astronomy_software?id=mvf-to-ms-configuration

These image products, interim or final, may be stored on the system as needed for active scientific analysis. Projects manage their open data release and distribution processes.

We estimate typical values of the data volumes for a typical observation using the 32k correlator mode over $T_{obs} = 8\,\mathrm{hours}$, with all the available correlation products (HH, VV, HV, VH), all 64 antennas, and an 8 second dump time, processed with the IDIA pipeline according to Figure 2, for the example use case of the MIGHTEE survey. For smaller data sets (e.g. observed in 4k correlator mode), the volumes will scale according to Equation 3.

These observing parameters produce a raw data set D1 that is 12 TiB in volume. If we apply the static RFI mask on D1 data-product as described in section 3.2, the resulting data is further split into the D2.1 data-products: spectral-line (excludes everything above 1420 MHz), continuum data is averaged by a factor of 32. The total size of the D2.2 data products is approximately 9 TiB, which includes the *data*, *model* and *corrected* columns. After calibration, only the corrected data products remain, for both the unaveraged (spectral-line) and averaged (continuum) data products, leaving calibrated D3 products that are approximately 3 TiB in size. The final images (spectral-line, polarisation and continuum cubes) and the calibration tables that comprise D4 are all created (or counted) at the end of the entire process, and are approximately 1 TiB in volume.

The reduction in volume of data sets from removal of the persistent RFI bands discussed in section 3.2 are not included in these estimates. Adding this for the spectral line and polarisation data sets will reduce the volume of data by $\sim$30%. We note that there are many factors that will affect the sizes of the various data sets (D1...4), which will affect the rate at which the storage volume increases, and the resulting storage pressure.

5 Resources and Operations

To explore the implication of a continuous stream of MeerKAT project observations, we have developed a mode to simulate the end-to-end process – from observations to science images. The model is based on the processing and data flow outlined in section 4. The outcomes of that study inform the data management policy outline in section 4.3. More detailed documentation on this model is available upon request.

MeerKAT projects may retain D2.1 data products for a limited time during the initial stages of the processing and scientific QA to allow for flexible, easy development and refinement of processing workflows. However, on a time scale of several months it is expected that processing workflows

will mature to a point where D2.1 data sets can be deleted as soon as D3 products (calibrated visibilities) are validated. D3 data sets may be retained for the duration of the project.

Regular observations will produce a regular stream of data to Ilifu, which will need to transferred, stored and calibrated effectively to ensure that projects can process their data in a reasonable time frame. The operations plan must provide ongoing storage and compute capacity to support the imaging projects as well as the broader community of users.

From a science perspective, we note that there is also a balance to be struck between retaining D2.1 data products and an initial QA window. It certainly is possible, at least initially, to retain D2.1 data while pipelines and workflows are being developed and optimised, but there is a threshold beyond which the retention of D2.1 data cripples the operation of the cluster as a whole. Optimising pipelines will make a profoundly positive impact to both project and Ilifu operations – raw data products could be retained for a shorter period, thus alleviating storage pressure.