Tricolour: an optimized SumThreshold flagger for MeerKAT

Benjamin V. Hugo\textsuperscript{1,2}, Simon Perkins\textsuperscript{1}, Bruce Merry\textsuperscript{1}, Tom Mauch\textsuperscript{1} and Oleg M. Smirnov\textsuperscript{2,1}

\textsuperscript{1}South African Radio Astronomy Observatory, Cape Town, South Africa; bhugo@ska.ac.za

\textsuperscript{2}Rhodes University, Makhanda (Grahamstown), Eastern Cape, South Africa

Abstract. We present Tricolour, a package for Radio Frequency Interference mitigation of wideband finely channelized MeerKAT correlation data. The MeerKAT passband is heavily affected by interference from satellite, mobile, aircraft and terrestrial transponders. Coupled with typical data rates in excess of 100 GiB/hr at 208kHz channelization resolution, mitigation poses a significant processing challenge. Our flagger is highly configurable, parallel and optimized, employing Dask and Numba technologies to implement the widely used SumThreshold and MAD interference detection algorithms. We find that typical 208kHz channelized datasets can be processed at rates in excess of 400 GiB/hr for a typical L-band flagging strategy on a modern dual-socket Intel Xeon server.

1. Radio Frequency Interference impacts on MeerKAT observation

Radio Frequency Interference (RFI) can be defined as any interfering signal that negatively affects the observable instrumental bandwidth of a radio telescope. The most prominent RFI impacting cm wavelength observations using MeerKAT (Jonas et al. 2018) are spatially coherent man-made telecommunications and navigational satellite transmissions. These substantially impacts all observation, particularly that of L-band, shown in Fig. 1 for a typical observation.

Figure 1. Mean visibility amplitude of a typical unflagged and uncalibrated L-band MeerKAT observation. Darkened areas show subbands where persistent interference is most likely, and serves as a “static” background mask during application of mitigation routines.
RFI affecting large portions of the bandwidth presents a substantial risk for, especially, sensitive continuum radio science projects, because the sensitivity of such observations scale as the square root of the observable bandwidth. As shown in Fig. 1, man-made interference power is orders of magnitude above the relatively dominating sub-Jy Active Galactic Nuclei (AGN) population at cm wavelengths — even on a bright calibrator field! A large number of science cases require very long observations to target imaging sensitivities at the level of only a few µJy/beam RMS noise, e.g. Mauch et al. (2020). Multi-hour observations using all available bandwidth is only one factor determining image observation sensitivity — another important factor is the number of unique interferometer spacings (baselines) contributing to the final synthesized image. This scales, roughly, quadratically with the number of antennae in the interferometric array.

The dump rate of MeerKAT’s raw correlated visibilities (excluding metadata) is 0.14 and 1.10 TiB/hr for coarse and fine channelization respectively, at a typical dump rate of 8s. RFI mitigation on a large wideband sensitive array therefore poses a significant processing challenge, requiring a flexible and parallel implementation of mitigation routines.

2. RFI detection and mitigation

Outlier detection routines are commonly used to detect and mitigate RFI in interferometry. The results are stored in a spectro-polarimetric flag array with the same shape as the data column, which is used in further calibration and imaging of the radio product.

Mitigation of man-made RFI is substantially simplified by two properties of the measured signal. Firstly, transmission signals are ordinarily highly polarized. Much of the bright cm dominating radio AGN population is linearly polarized to a couple of percent. Secondly, a radio interferometer measures phase information of the spatial locality of emitting sources. The complex vector rotates quickly on long spacings for all declinations away from the celestial poles. Through complex averaging, a substantial portion of the detectable RFI is washed out.

Various simple and effective outlier detection techniques that can be used to detect the remaining RFI not washed out in the fringes of the interferometer, are discussed in Offringa et al. (2010). The SUM_THRESHOLD algorithm is a simple and fast windowed procedure widely used for RFI detection that can be applied to two dimensions — that of time and frequency per spacing. The method clips values where the sum of the values within a time and channel window exceeds an interatively adjusted average threshold. The method is very sensitive to RFI while maintaining good false-positive classification rates for large enough window sizes. The algorithm is trivially parallel over the unique interferometer spacings (baselines), but requires a reordering step to select the time and frequency subsets per baseline from data that is traditionally stored in time-sorted row order. For a large array such as MeerKAT this step is amortized by the dominating run time of the sum-thresholding for various window sizes and number of iterations. To aid in background estimation on short spacings, we provide options to apply “static” masks to flag large bands of persistent RFI (such as the one shown in Fig. 1) on adjustable ranges of baseline length, as well as options to form residual products based on predicted sky models.

The sizes of the windows, as well as the sensitivity adjustment parameter, have to be fine tuned by hand for the particular science case and channelization regime of
Tricolour: an optimized SumThreshold flagger for MeerKAT

Figure 2. Here we show bandwidth and spatial flagging statistics after applying typical multi-pass flagging strategies for L (first row) and UHF (second row) bands of calibrator PKS B1934-638. UHF band is generally very clean compared to L-band, however we note that the large spread in visibility amplitudes at the bottom of the band is due to the substantial cumulative contribution from the off-axis AGN population, due to the wide field of view of the instrument. The contribution from navigational systems is substantially reduced in comparison to what was seen in Fig. 1.

the observation, to minimize false positives. To achieve this we provide a YAML-based interface to the user to allow users to customize strategies to their requirements. Typical flagging statistics and results for MeerKAT are shown in Figure 2.

3. A parallel, scalable implementation

Our flagger is distributed as an Open Source Python package, available on the Python Package Index as “tricolour”. We have implemented the SumThreshold method written in Numpy (Harris et al. 2020) and cache-optimized Numba (Lam et al. 2015) kernels. The user-available version of the flagger ingests data from Measurement Set v2.0 columns exposed as chunked Dask Arrays by Dask-ms. This architecture is more fully described in Perkins et al. (2021a). A trimmed version of the flagger is used in the online MeerKAT Science Data Processor which flags most of the substantial RFI in buffered raw data at line rates, before storage to the MeerKAT archive.

We profiled our implementation with a dual-socket Intel Xeon 8160 system with 550 GiB of memory applying a typical multi-pass L-band flagging strategy to a sufficiently large (> 100 GiB) coarsely channelized (208kHz) dataset captured at a dump rate of 1s. After tuning thread affinity and chunk size to optimize cache performance (through non-intrusive L3 cache profiling with the Linux kernel utility perf) we show that our flagger scales well to more than 20 physical cores (Figure 3) and flags data at over 400 GiB/hr.

4. Conclusions

We have built a highly configurable and scalable RFI flagger for the MeerKAT radio telescope, that is capable of processing coarsely channelized wideband data well above
line rate with typical flagging strategies. Our flagging software has substantially cut down on the processing time needed for telescope commissioning, as well as recent continuum science observation processing through the CARACa (Jozsa et al. 2021) pipeline, as well as processing of the MIGHTEE continuum survey (Delhaize et al. 2020) through the Oxkat (Heywood 2020) pipeline.

While Tricolour is currently restricted to operation on a single node, Dask offers the potential for horizontal scaling on a cluster, as described in Perkins et al. (2021b). As we support data ingest through the generic, defacto standard, Measurement Set v2.0 interface, modification of the flagger to support instruments other than MeerKAT should not be difficult to accomplish.

Acknowledgements

All figures in this paper were generated using Matplotlib - a 2D graphics package used for Python for application development, interactive scripting, and publication-quality image generation across user interfaces and operating systems. The MeerKAT radio telescope is a facility operated by the South African Radio Astronomy Observatory. A business unit of the National Research Foundation of South Africa.

References

Delhaize, J., et al. 2020, MIGHTEE: Are giant radio galaxies more common than we thought? MNRAS, accepted
Harris, C. R., et al. 2020, Nature, 585, 357
Heywood, I. 2020, oxkat: Semi-automated imaging of MeerKAT observations. 2009.003
Jonas, J., et al. 2018, in MeerKAT Science: On the Pathway to the SKA (SISSA Medialab), vol. 277, 001
Jozsa, G., et al. 2021, in ADASS XXX, edited by J.-E. Ruiz, & F. Pierfederici (San Francisco: ASP), vol. TBD of ASP Conf. Ser., 999 TBD
Lam, S. K., et al. 2015, in Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC, 1
Mauch, T., et al. 2020, The Astrophysical Journal, 888, 61
Offringa, A., et al. 2010, Monthly Notices of the Royal Astronomical Society, 405, 155
Perkins, S., et al. 2021a, in ADASS XXX, edited by J.-E. Ruiz, & F. Pierfederici (San Francisco: ASP), vol. TBD of ASP Conf. Ser., 999 TBD
— 2021b, in ADASS XXX, edited by J.-E. Ruiz, & F. Pierfederici (San Francisco: ASP), vol. TBD of ASP Conf. Ser., 999 TBD
This figure "mkLRFI.png" is available in "png" format from:

http://arxiv.org/ps/2206.09179v1
This figure "mk_lband_1934_band.png" is available in "png" format from:

http://arxiv.org/ps/2206.09179v1
This figure "mk_lband_bandpercentages.png" is available in "png" format from:

http://arxiv.org/ps/2206.09179v1
This figure "mk_lband_uvpercentages.png" is available in "png" format from:

http://arxiv.org/ps/2206.09179v1
This figure "mk_uband_1934_band.png" is available in "png" format from:

http://arxiv.org/ps/2206.09179v1
This figure "mk_uband_bandpercentages.png" is available in "png" format from:

http://arxiv.org/ps/2206.09179v1
This figure "mk_uband_uvpercentages.png" is available in "png" format from:

http://arxiv.org/ps/2206.09179v1
This figure "tricol_occ.png" is available in "png" format from:

http://arxiv.org/ps/2206.09179v1
This figure "tricol_scaling.png" is available in "png" format from:

http://arxiv.org/ps/2206.09179v1