HERA Memo 71
On Data Rates: Saving Short Baselines
August 9, 2019

C.L. Carilli[1,2], B. Nikolic[2], N. Thyagarajan[1], P. La Plante[3], J. Kent[1]

ccarilli@aoc.nrao.edu

ABSTRACT

We argue that saving the raw data on short baselines is both imperative for the project, and within the scope of the HERA data rate out of the Karoo. Transferring all visibilities on all baselines shorter than 50 m requires 10% of the total bandwidth out of the Karoo currently allocated to HERA, under very conservative assumptions as to time and frequency averaging, and useful observing hours per day.

1. Why save the raw data?

A guiding principle in the HERA software development has been generality and reproducibility by third parties. The most intensive processing of the data, in terms of determining data quality and reliability, and likely the dominant factor in determining the final power spectra results, is the real time system. This system involves flagging, calibration, and substantial averaging over different axes, including potentially LST and redundant triad binning (La Plante et al. 2018, HERA Memo 57). If only the averaged, binned, flagged, and calibrated data are saved, and the raw data is deleted, this puts tremendous pressure on the RTP to get it right, since there is no going back. Moreover, this process then violates the 'reproducibility' requirement of the software, since the RTP processing cannot be reproduced if the raw data is deleted. As HERA approaches its sensitivity limits, intepreting results at the 10^{-5} level will invariable reveal unexpected effects, and it will be impossible to determine if the effects are the result of the (non-reproducible) RTP, or post-processing.

[1]National Radio Astronomy Observatory, P. O. Box 0, Socorro, NM 87801

[2]Cavendish laboratory, Cambridge University, UK

[2]Univ. Pennsylvania, Philadelphia, PA

A simple example would be if a single bad day of data is added during the LST binning process. This would ruin an entire season of observations, if the raw data is deleted. When pushing to very low levels, 'bad' may not be identifiable in real-time, but only through detailed post-analysis of all the data. Similarly, we are already aware of significant non-redundancies in the array (Kent et al. 2018, HERA memo 55; Kent et al. in prep.; Fagnoni et al. 2019, submitted), and hence binning of redudant baselines automatically leads to imperfections. Both LST and redundant triad binning preclude subsequent antenna-base self-calibration.

Lastly, the closure phase delay spectrum approach for power spectral estimation requires data in which the closure relations have not been violated through potentially non-linear processing related to the combination of calibration, averaging, and binning. The closure phase analysis requires data as close as possible to raw.

On the other hand, there is a major data rate issue looming.

In this memo, we propose that at least some of the 'raw data'[1] is saved, namely, the shorter baselines. The shorter baselines of HERA are currently the focus of the power spectral search using the delay spectrum approach. With suitable averaging, and careful selection of optimal time ranges, the raw data rates can be reduced to a level that is within the scope of the current HERA allocation out of the Karoo. While saving the short baseline data will not allow for sky calibration using the out-riggers, it may allow for a re-analysis of redundant calibration, and, if the other sky-based calibration terms are saved each day (such as sky-based bandpasses using the longer baselines), then it may allow for a semi-recalibration of a season's data.

2. Data Rate

Following are the definitions and assumptions for the data rate calculation.

The definition of Terabyte is: 1 TB = 1024 x 1 GB = 1e12 Bytes = 8e12 bits. The correlator produces 64 bits, N_{bits}, per individual complex visibility (32 bits for Real, and 32 for Imaginary).

To determine the number of baselines, N_{base}, out to a given length, a mock single-record

[1]Raw data here is defined as output from the correlator that has been averaged in time and frequency, but no calibration applied, nor any LST or triad binning. Some flagging could be performed before averaging, but how flagging is implemented needs to be studied.

observation was generated for the HERA351 split core configuration (Dillon et al. 2016, ApJ, 826, 181), using SIMOBSERVE in CASA. The CASA plotms tool then has simple procedures to determine the number of baselines as a function of maximum length. These values are shown in Table 1.

For the channel width, we adopt the same resolution as is now employed in HERA50 of 100 kHz. The number of channels from 50 MHz to 250 MHz is then, $N_{ch} = 2048$. We adopt an averaging time, or record length, $\Delta t = 16$ s, on all saved baselines (currently the default in BDA for the short baselines). These assumptions are easily adequate to assure there is little loss of coherence on the baselines in question (see appendix). In fact, these could be increased by factors of a few without loss. However, maintaining some time and frequency resolution remains important for flagging, in particular, if no real time flagging is performed prior to averaging.

We assume that there are 6 hours of quality observing time (t_{obs}), each night. For the current calculation, we assume only 2 polarizations are saved (EE, NN; $N_{pol} = 2$), and that the differenced data is not saved. These criteria are the minimum amount of data required to employ the closure phase delay spectrum approach.

Real-time flagging of the visibilities at higher frequency and time resolution using a simple thresholding process might be employed prior to averaging and transmission, with the caveat that, while threshold flagging in frequency on a record by record basis can be implemented in real-time, thresholding in time may be problematic in the RTP, since it requires buffering. At the minimum, flagging in real-time could employ a look-up table of known bad channels, such as Orbcom and FM radio. Further investigation is required to determine how flagging at narrow channels, then averaging, affects the linearity of the visibilities, in terms of maintaining closure relationships and related. Of course, one could adopt a conservative procedure that if one channel is bad, the full output channel is flagged.

The implied data volume per day is then:

$$Volume = N_{base} \times N_{pol} \times N_{ch} \times (t_{obs}/\Delta t) \times N_{bits} \quad \text{bits}$$

Table 1 shows the data rates. The raw data (after averaging), for all baselines shorter than 50 m sums to 0.21 TB per day. The total data rate out of the Karoo currently allocated to HERA is 2 TB per day. Hence, saving the raw data for baselines out to 50 m requires 10% of the HERA bandwidth.[2]

[2]There is roughly a factor two difference between the bottoms-up calculation of the data rate presented

Table 1: Data Rates

Max. Baseline Length	15 m	30 m	50 m	100 m
Number of Baselines	839	2377	4901	17983
Data Rate (TB/day)	0.04	0.11	0.22	0.80

If we want to save all polarizations, and the differenced data, the data rate would increase by a factor of three. Hence, further reductions may be required, such as de-selecting known bad frequency ranges, such as those around the FM band and Orbcom. Also, the averaging time for baselines shorter than 100 m could be increased substantially (factors of a few), with no loss of coherence (see Appendix). Likewise for the channel width, although maintaining spectral resolution allows for exploration of higher k-modes in the power spectrum. And there may be an increase in the allowed HERA bandwidth out of the Karoo, in which case this is not an issue.

3. Requirements

We propose saving the short baseline data each day. The requirements for the system are:

- Possible initial processing involving threshold-based flagging of the full resolution visibilities out of the correlator. This step may piggy-back on the current RTP, depending on implementation, although it is not absolutely necessary, if implementation is difficult.
- Averaging in frequency and time.
- Local storage for the data each day (of order 0.2 TB).
- Spigut for transmission out of the Karoo.

herein (meaning, starting from bits per visibility and number of visibilities), and the top-down calculation in the current jupyter notebook calculator (meaning, scaling the maximum date rate down by various averaging and selection criteria). The cause for this difference is under investigation, but, again, the results are within a factor of two, which can be accommodated with further averaging in time, if necessary.

4. Appendix: a simple consideration of time and bandwidth smearing

Time and band width smearing are well covered in text books (eg. TMS, SIRA, Condon & Ransom). We present a useful approximation of the effects.

Bandwidth smearing (chromatic aberration): for a channel width, $\Delta\nu$ at an observing frequency, ν, the radially smeared source size, $\Delta\theta_{\Delta\nu}$, is given approximately by:

$$\Delta\theta_{\Delta\nu} \sim (\Delta\nu/\nu) \times \theta_{PC}$$

where θ_{PC} is the distance of the source from the phase tracking center (assumed to be the antenna pointing center). Assume a source is at the edge of the primary beam of an antenna of diameter, D, or at a distance from the phase center $\sim$ FWHM of the primary beam: $\theta_{PC} \sim \lambda/D$.

The scale of the fringe pattern, or 'synthesized beam', for a baseline of length, B, is $\sim \lambda/B$. We want the smeared source size to be some small fraction, X, of the synthesized beam size, leading to:

$$X \times (\lambda/B) = \Delta\theta_{\Delta\nu} = (\Delta\nu/\nu) \times (\lambda/D)$$

Hence, in terms of fraction of the synthesized beam size, the smeared source size near the edge of the primary beam, becomes:

$$X \sim (\Delta\nu/\nu) \times (B/D)$$

For $\delta\nu = 0.1$ MHz, $\nu = 50$ MHz, $D = 15$ m, and $B = 50$ m, the smearing size, $X \sim 0.7\%$ times the size of the synthesized beam. For a 15 m baseline, this reduces to $\sim 0.2\%$.

These values are approximate, but within a factor two of the coherence of a visibility on a given baseline. In either case, bandwidth smearing is not an issue for HERA short baselines, for a channel width of 100 kHz.

Time smearing: for a record length, Δt, the azimuthally smeared source size, $\Delta\theta_{\Delta t}$, is given approximately by:

$$\Delta\theta_{\Delta t} \sim (\Delta t/24\ hrs) \times \theta_{PC}$$

where, again, θ_{PC} is the distance of the source from the phase tracking center. Analagous

to the calculation above, for a source at the edge of the synthesized beam for an antenna of diameter D, and baseline of length B, the smeared source size will be some fraction of the synthesized beam size, X, as dictated by:

$$X \times (\lambda/B) = \Delta\theta_{\Delta t} = (\Delta t/86400\ s) \times (\lambda/D)$$

The smeared source size near the edge of the primary beam, in terms of a fraction of the synthesized beam, is then:

$$X \sim (\Delta t/86400\ s) \times (B/D)$$

For $\delta t = 16$ s and and $B = 50$ m, the smearing size is $\sim 0.06\%$ the size of the synthesized beam. For a 15 m baseline, this reduces to $\sim 0.02\%$.

While approximate, it also appears time smearing is not an issue for HERA, even for record lengths significantly longer than 16 sec, for baselines shorter that a few hundred meters.