LWA Data Management

Authors

Greg Taylor	(UNM)
Jayce Dowell	(UNM)
Ylva Pihlstrom	(UNM)
Joe Craig	(UNM)

December 3, 2012
LWA Memo #177

Contents

1 Introduction

The first LWA station, LWA1, has now been in operation for a year. This memo describes the mechanisms by which users of the LWA1 will retrieve their data for subsequent analysis. While there has been some discussion of how to move data from stations to the correlator (Taylor 2007, Taylor & Ray 2008), prior to this memo nothing has been written about getting data to users after it is recorded at the site. Given that there is useful science that we can do with LWA1 we need to establish the mechanism by which users can collect their data.

Since LWA1 is capable of capturing large amounts of data, users will have to select between a limited number of options for data retrieval. And since the LWA1 budget is constrained, some of the choices may involve modest effort and expense on the part of the user. By way of examples in this memo we assume that users are interested in the widest bandwidth beamformed DRX output which produces 602.8 Mbps (19.6 MSPS, 4 bit I + 4 bit Q, 2 tunings, 2 polarizations), or the TBN mode which has a slightly higher data rate, 812.1 Mbps (100 KSPS, 8 bit I + 8 bit Q, 2 polarizations), at maximum.

We also describe some options related to the newly deployed LWA1 User Computing Facility (hereafter UCF) which is located at the site. Current wisdom suggests that it is generally more efficient to move the computing facilities close to the data rather than moving the data to the computing facilities. Also, not every LWA1 user may have facilities capable of handling the 20+ TB of data produced by a typical LWA1 observing program.

We furthermore assume that data can be physically collected from the site only once/two weeks by the UNM project office. This may change but is likely to be the case indefinitely. Users who desire to go out to the site for their observations are welcome to propose a more frequent collection of data.

2 Spectrometer Mode

Most users employing beams should be satisfied with the new (Nov. 2012) spectrometer observing mode which provides Stokes I, Q, U, and V for 1024 channels at a time resolution of 40 milliseconds with the widest bandwidth mode (filter 7, 19.6 MSPS). Some trade-off between integration time and number of channels is possible up to a maximum of 2048 channels at 160 millisec and down

Table 1. Integration Times for DRX Filter #7

Channels	Integration Count				
	384	768	1536	3072	6144
64	0.001	0.003	0.005	0.010	0.020
128	0.003	0.005	0.010	0.020	0.040
256	0.005	0.010	0.020	0.040	0.080
512	0.010	0.020	0.040	0.080	0.160
1024	0.020	0.040	0.080	0.160	0.321
2048	0.040	0.080	0.160	0.321	No
4096	0.080	0.160	0.321	No	No

to 256 channels at 10 millisec (see Table 1). The spectral resolution and integration time can be determined from:

$$channelwidth = samplerate/transformlength$$

and

$$integration = transformlength * integrationcount/samplerate$$

where *samplerate* is in samples/second, *channelwidth* is in Hz, $transformlength$ is the number of channels, and *integrationcount* is an integer. By way of example consider the widest DRX bandwidth sample rate (19.6×10^6 samples/sec filter code 7), *intgrationcount* of 768, and 1024 channels gives a channel width of 19.140 KHz and an integration time of 40 milliseconds. The file size for this typical spectrometer observation (1024 channels, 40 millisecond integrations) for 1 hour is 2.74 GB at a data rate of 6.2 Mbps. All such observations will be archived and available for download from:

`http://lda10g.alliance.unm.edu/archive/list.py`

In the case of spectrometer data, no data management plan is necessary in the proposal.

3 The Raw Data Option

For all TBW/TBN observers and for observers who require no averaging of the DRX beam output in time or frequency (e.g., pulsar observations, certain solar observations) it will be possible to request the raw data. Justification for this choice will be required in the observing proposal. Furthermore all observations planned using the raw data option will require a data management plan that specifically describes how the user will cope with the large data volumes generated. The specific issues that should be addressed are:

1. an estimate of the total data volume,
2. where the user plans to reduce the data (UCF, UNM, home institute)

3a. If UCF/UNM: how much time they need to crunch the data and retrieve the results.

3b. If at home institute: how they plan to get the data back to their home institution, e.g. how many drives they will need.

By way of example there are a couple possibilities including **(A)** Reduce the data at the User Computing Facility. The UCF (see Dowell 2012) is a cluster of 6 hexacore nodes each equipped with two CUDA compatible GPUs. The UCF is located in the old VLA correlator room within the VLA control building thanks to the generosity of NRAO for hosting it, and JPL, VT and UNM for supplying the nodes. The cluster is connected to LWA1 by a dedicated fiber so that it is possibly for the LWA1 operator to copy raw data soon after the observations to the cluster. The cluster has a common storage area currently limited to 22 TB, corresponding to roughly 60 beam-hours. Raw data will be deposited to the commons by the operator. Users should process the raw data from the common area and inform the operator when the raw data may be deleted. Each node has 3 TB of local storage, and more local storage can be accomodated as needed. **In general the UCF time will be granted immediately following the observations.** If there is a time-window during which the user cannot make use of the cluster in a timely fashion then scheduling constraints for the observations should be provided. All raw data will be archived in a default channelization and time averaging and full Stokes.

(B) Ask for data to be processed at UNM. A limited amount of data can be brought back from the site and processed at UNM. This will require a UNM collaborator and proper coordination to be described in the proposal.

(C) Ask for data to be shipped to the user. Valid options for storage media may be either (1) properly tested and configured DRSUs (Data Recorder Storage Units; Wolfe, Ellingson & Patterson 2009); or (2) external USB hard drives of size 2 TB or larger formatted as ext2/3/4, or bare drives of size 2 TB or larger; or (3) by transferring raw data over the internet (requires a very good connection). For CFP2 and the duration of the URO there are limited funds set aside for the purchase of external drives so users need not supply these. In the case of recording onto user-supplied DRSUs the units will be installed at the site and then collected after observations complete (but not less than two weeks). In this case only the data from that user will be written onto the DRSU.

4 MCS Metadata

MCS-generated metadata for all projects will be deposited at a location from which they can readily retrieved, currently:

`http://lda10g.alliance.unm.edu/metadata/observation/`

5 Proprietary Period

All LWA users will have a one-year proprietary period, after which data in the LWA archive will be made available to anyone requesting it. Data obtained from the archive after the proprietary period has passed will not be subject to the LWA publication policy except for the requirement that they include the following statement in their acknowledgements: "Support for operations and continuing development of the LWA1 is provided by the National Science Foundation under grant AST-1139974 of the University Radio Observatory program." The current archive distribution scheme provides some limited protection from outside access by unregistered users, but registered users will be able to see all files. For now we ask that users only download their own data from the archive.

6 Summary

The recent availability of a spectrometer mode with good spectral resolution should dramatically reduce the data volume from LWA1 during routine observations. For those still wanting access to the raw data then this will be possible but users will need to describe in a data management plan how they intend to cope with the large data volume. Some possibilities include, but are not limited to, use of the newly installed User Computing Facility, and data shipment. We have described methods by which we will manage this activity that will require some action from LWA users. We will learn as we go about exactly what data rates can be supported by the UNM network.

This plan describes LWA1 data management activities while LWA1 is supported as a University Radio Observatory which is expected to extend through 3/1/2015. Depending on the availability of funding, these services may change.

References

Dowell, J. 2012, LWA Memo #193

Taylor, G. B. 2007, LWA Memo #110

Taylor, G. B., & Ray, P. 2008, LWA Memo #131

Wolfe, C., Ellingson, S., & Patterson, C. 2009 LWA MCS memo #0019, `http://www.ece.vt.edu/swe/lwavt/doc/MCS0019.pdf`