LWA Data Management

Authors

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

February 11, 2011
LWA Memo #177

Contents

1 Introduction

The first LWA station, LWA1, is soon to commence operations. This memo describes the mechanisms by which users of the LWA 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 severely constrained, some of the choices may involve effort and expense on the part of the user. By way of examples in this memo we assume that users are interested in the beamformed DRX output which produces 627 Mbps (19.6 MSPS, 4bitI+4bitQ, 2 tunings, 2 polarizations). The TBW mode and TBN modes have slightly higher data rates, approaching 1000 Mbps.

We furthermore assume that data can be physically collected from the site only once/week by the UNM project office. This may change in the future but is likely to be the case for much of 2011. Users who desire to go out to the site for their observations are welcome to propose a more frequent collection of data.

2 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. In this case it will be up to the user to provide the storage medium at least one week prior to the commencement of observations. 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. A recommended brand is the Iomega Prestige 2 TB. Other drives are allowed but should be of similar dimensions. USB powered drives are not allowed.

In the case of recording onto user-supplied DRSUs the units will be installed at the site and then collected after one week. In this case only the data from that user will be written onto the DRSU.

Given the storage capacity of the DRSUs at 5 TB each and two DRSUs allowed per MCS data recorder, we arrive at 6.761 TB/day or about 36 hours until both DRSUs are filled up. This time may be spread throughout the 168 hours in the week in any way desired, but the average duty cycle will be less than or equal to 21%.

In the case of recording onto user-supplied external USB drives, the maximum number of drives allowed is 5, so assuming 2 TB drives, one reaches the same duty cycle as above for DRSUs. An added complication is that copying data from the DRSUs to the USB drives may require halting the MCS executive. This copy process will likely happen at about half real-time. So if 36 hours of data are collected, it will take 72 hours to copy the contents over to the external drives. The net duty cycle remains the same as with DRSUs at less than or equal to 21%.

3 The Processed Data Option

For some LWA1 projects it may be possible to average up the data in time and/or frequency sufficiently such that the data can be archived at UNM and delivered over the internet to the user. Data will be written in FITS format by the LSL package.

Tests of internet connections from UNM to Indiana and Stanford have revealed connection rates of between 8 and 30 Mbps. Since real time is not required we figure that an acceptable average data rate is 20 Mbps. Thus the raw DRX data needs to be averaged by at least a factor of 30, or LWA capture modes requiring less bandwidth (e.g., with a single tuning and polarization) averaged by a lesser factor to achieve a reduction to 20 Mbps. Averaging by this factor 30 would still allow 4096 channels and an integration time of 0.01 seconds. As perhaps a more "typical" example we considered averaging to 4096 channels and 1 second integrations. Using 64 bits/channel gives a data rate of 0.26 Mbps.

The data processing will be performed using LWA data recording computers (one per beam), during times when they are not actively recording data. Tests on simulated data with these computers indicated that the processing is in the worst case 1/3 of real time. Assuming that data can be copied to an internal drive for processing at 1/2 of real time, our current estimate is that it will take a factor of 6 longer than real time to process the data. This is conservative and may be somewhat faster for lower data rates. The duty cycle in this conservative case is about 14%. But there is the significant advantage that data can be retrieved simply by accessing the LWA archive at

`http://fornax.phys.unm.edu/lwa/archive/list.py`

The total amount of data delivered in the most conservative case is 324 GB/week and could be much less (only 3 GB for 1 second averaging and 4096 channels).

We are not completely confident that the UNM network will be up to the task of serving up 300+ GB/week to multiple users, but it is not clear that this will acutally be needed. If it is then we may have to resort to requesting disks from users for data transfer, and/or requesting users to increase the averaging time to at least 0.1 seconds.

4 MCS Metadata

MCS-generated metadata for all projects will be deposited in the LWA archive and may be retrieved from: `http://fornax.phys.unm.edu/lwa/archive/list.py`

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. 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

Keeping up with LWA1 data rates using all 4 beams and TBN/TBW presents a challenge. We have described methods by which we will manage this activity that will require some action from LWA users. Users will need to select their preferred option for data retrieval, and will either need to supply disk drives, or supply parameters over which their data may be averaged in frequency and time. Initially this information will be manually requested using a simple form. In the near future we plan to develop software that will assist users in the planning of their observations including specifying options for data retrieval. 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 prior to 9/30/2011 when the ONR contract expires. Depending on the availability of funding, these services may change.

References

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

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

Wolfe, C., Ellingson, S., & Patterson, C. 2009 LWA memo #165