Initial specifications of xNTD pipelines

T.J. Cornwell, ATNF

Tim.Cornwell@csiro.au

20/9/06

Abstract: *As discussed by Johnston (2006), we expect that the first major scientific observations made with the xNTD will be all-sky continuum survey, monitor, and HI survey. Since the necessary processing is well understood, we are in a position to describe in some detail the necessary pipelines. We do so with the goal of being able to then derive the necessary software and hardware capabilities required to support the pipelines.*

1. Background

Initial estimates of the spectral line processing load for xNTD (Cornwell, 2005) showed that for the full spectral resolution, many thousands of current day processors would be required. Since it seems clear that the future advances in processor capability will come mostly from cost of and number of cores rather than from individual processor capability, we can expect that for 2011 many thousands of processors/cores will be needed. The data storage required for the spectral processing is many Terabytes per hour.

An immediate and important conclusion is that the current model of exporting the data to the desktop will not work – we don't expect either the network throughput or the desktop power to be sufficient. Hence all calibration and imaging must be done close to the telescope. Furthermore, the data volumes require that processing be conducted as soon as possible after the observations, in real time if possible, and then the observed data and derived images will have to be discarded and only the catalog retained.

In this memo, I go one level deeper than the previous memo to investigate the details of the pipeline processing to be performed on survey observations with the xNTD.

2. Initial analysis

For the proposed configuration of the xNTD (Johnston 2006, Cornwell, 2006), the array will be confusion limited in continuum within a total integration time of tens of minutes. To improve the Fourier plane coverage, this integration time should be spread over a wide range of hour angles. Hence an initial all sky survey of the roughly 1000 pointings required at 0.85GHz, 1.15GHz and 1.45GHz can be completed within a few weeks, say. At each frequency, this will provide an image of the entire sky, plus first, second, and

perhaps third order gradients in frequency. A model of the continuum sky will also be available as raw images and catalogs. The images will be defined for a set of standard field centers (SFC) chosen to cover the sky.

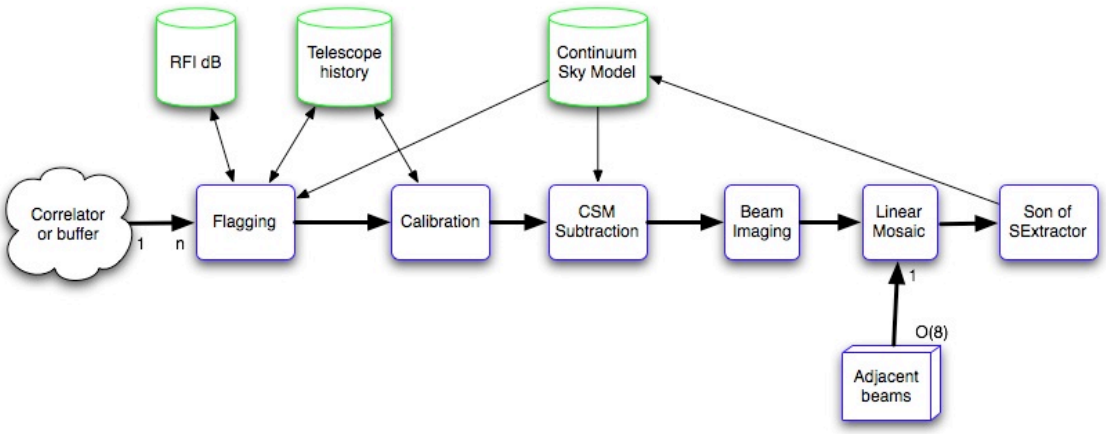
A deep HI all sky survey to z~0.2 will consume an entire year. Thus we can consider the two surveys separately and it makes sense to complete the initial continuum survey prior to embarking on the HI survey. This simplifies the processing steps considerably. For the continuum survey, we could choose to observe a given field for short periods of time, thus randomizing any systematic errors. In designing the processing for the HI survey, we can assume that the continuum sky is well known, except for any transients.

Continuum subtraction will be performed can be performed by calculating the visibility for each channel from the continuum model for a given SFC, either as image cubes or as components. The same mechanism can also be used for continuum calibration.

3. Continuum processing

For a continuum survey, the steps are roughly as follows:

1. Ingest data from correlator, reorganizing as necessary
2. Excise known RFI
3. Discover and flag bad data
4. Calibrate for various effects
5. Remove effects of known bright sources using best model of sky and telescope
6. Clean individual beams
7. Make linear mosaic image of adjacent beams
8. Search for brightest sources and add to catalog
9. Repeat steps 4 to 7
10. Archive results

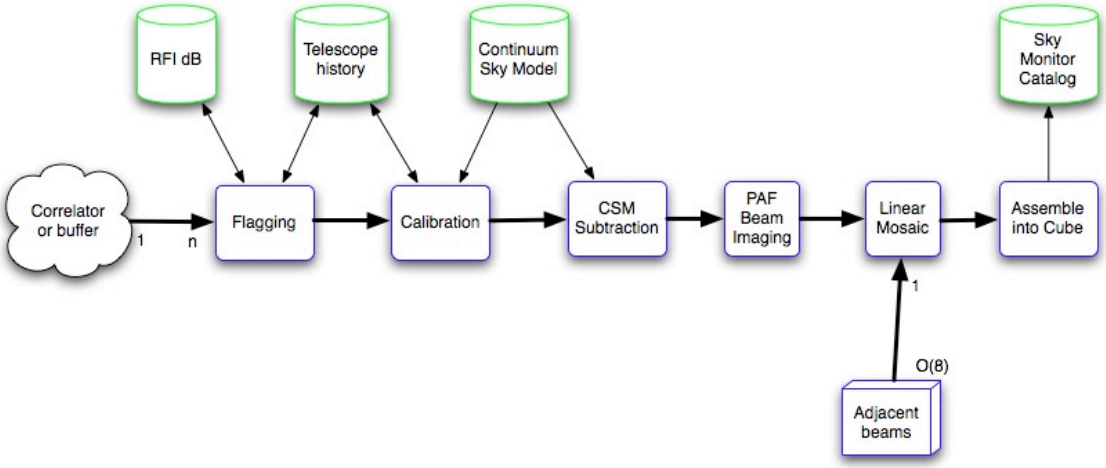


xNTD Continuum pipeline
- One per PAF beam
- Assumes beams fixed on sky
- Multiple pass to go deeper
- GSM subtraction has full model of telescope
- Requires data buffer

Figure 1 Pipeline for continuum processing

Timing the imaging step with AIPS++ running on the 64bit server *delphinus*, we find that cleaning one field to the noise level takes about 1600s – close to the observing time. With some optimization and more careful choice of parameters, we expect that this could be reduced by a factor between two and five. Hence it seems plausible that the entire continuum processing could run on one (2006) processor per PAF beam.

If run with a nominally correct sky model, the source extraction provides information on time variability. This then leads to a sky monitor capability:



xNTD Sky Monitor pipeline
- One per PAF beam
- Runs at rate allowed by resources allocated

Figure 2 Pipeline for sky monitor

4. HI pipeline

For an HI survey, the processing steps that must be performed are:

1. Ingest data from correlator, reorganizing as necessary
2. Excise known RFI
3. Discover and flag bad data
4. Calibrate for various effects
5. Remove continuum emission
6. Assemble mosaic cubes of adjacent PAF beams
7. Search for emission features
8. Archive results

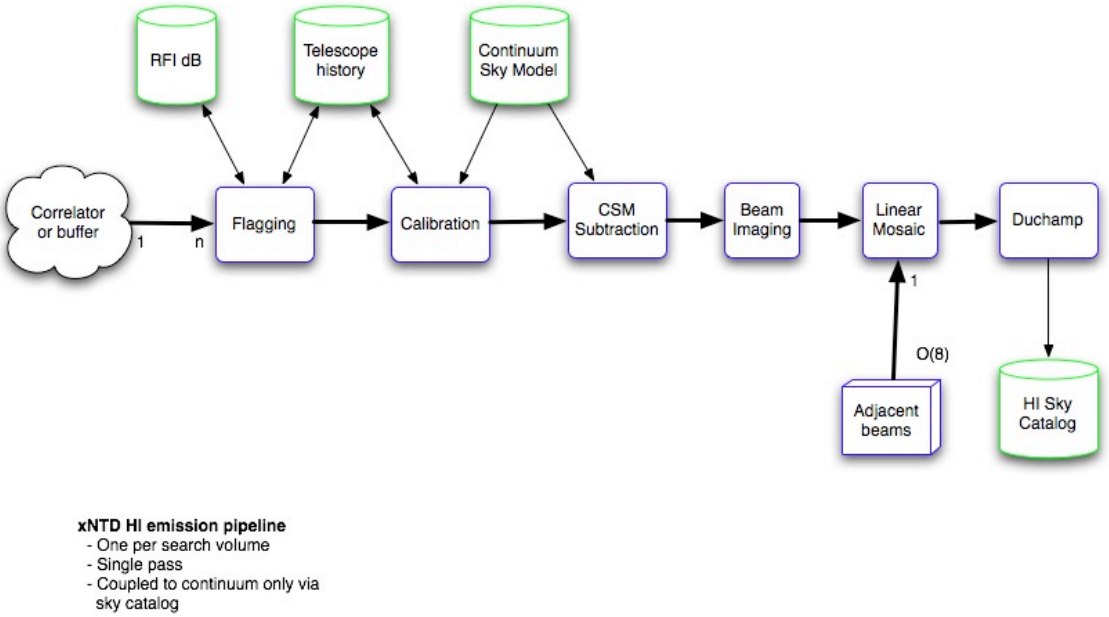


Figure 3 Pipeline for HI survey imaging

Here we make a fundamental assumption vital to the overall processing scheme that we can design xNTD such that most of the measurements made are independent per PAF beam and per channel. This, of course, is why we would choose to build a PAF-enabled telescope as opposed to an ATA or Big Gulp solution (Cornwell, 2005). If this assumption is correct then most of the spectral line processing can be partitioned per PAF beam and channel. The exception to this would be step 7, the search for structures in the spectral cube. If we don't desire to search for arbitrarily large structures in frequency, we can partition the search by windows in frequency space. The entire processing can then be partitioned by PAF beam and by search window.

Staveley-Smith (2006) has analyzed the optimum parameters for the HI survey. He proposes a minimum frequency resolution of 100kHz. Working from his numbers, and allowing a factor of two overlap in frequency, we find that the number of these search windows is 3840 (see Table 1). In this regime, the size of data per SV is of the order of a GB – well matched to an affordable memory size.

Hence we have a model for the HI survey processing – the data are partitioned according to SV and distributed to processors accordingly. The data streams into each processor over the entire observing time, being gridded onto a cube that fits entirely in memory. At the end of observing, the cube is Fourier transformed into the image plane. It must then be linearly mosaiced with the neighboring beams (requiring transfer of about 1GB to O(8) neighbors). The SV is then roughly 1GB.

Timing the beam imaging on *delphinus*, we have that one SV can be imaged in 3360s real time, 1600s CPU time. This part of the processing is therefore at least an order of magnitude faster than the observing, leaving good headroom for the rest of the processing.

The other substantial part of the processing is the subtraction of the continuum sky model from the data. This is unlikely to be more time-consuming than the imaging step.

Table 1 Array and observing parameters for Staveley-Smith's proposed HI survey

Number of antennas	30	
Diameter of antennas	12	m
Number of PAF beams	30	
Baseline length	4000	
Number of spectral channels	8192	
Number of continuum channels	32	
Upper frequency	1.8	GHz
Lower frequency	0.7	GHz
Bandwidth	0.3	GHz
Integration time	10	s
Spectral line integration time	12	h
Continuum integration time	0.5	h
Bytes per data point	32	4 pol, 8 bytes
Spectral line data rate per beam	23.5930	MB/s
Continuum data rate per beam	0.0922	MB/s
Spectral line data size per beam	1019.22	GB
Continuum data size per beam	0.17	GB
Pixels per beam	1333	
Spectral line image size per beam	58.25	GB
Continuum image size per beam	0.23	GB
Spectral search window	128	channels
Number of search volumes	3840	2* overlap
Data per search volume	15.93	GB
Image size per search volume	0.91	GB

5. Open questions

1. Incremental versus single shot observing. Johnston has proposed an incremental observing mode whereby each field is observed over and over every few days, thus building up the total integration over a period of a year. This would have considerable implications for the model proposed here.
2. Refactoring of pipelines – the actual work performed in the pipelines can probably be refactored to reduce the overall amount of work.
3. Bandpass calibration – the bandpasses can be calibrated separately per frequency block but overall continuity between the edges must be maintained.
4. How to deal with boundaries in SVs? Boundaries in the searches will lead to catalog defects unless some care is taken.

6. Summary

We have described an approach to xNTD pipeline processing that parallelizes well. The

angle-angle-frequency volume sampled by the array is divided into search volumes. A given search volume (SV) corresponds to given PAF beam, supplemented by the adjacent beams, and for the HI survey partitioned in frequency.

Using this approach, a continuum survey can be reduced by about 30 computational nodes (one per PAF beam), and a cluster of about four thousand computational nodes can process an HI survey of the parameters proposed by Staveley-Smith. Each node could be a typical 2006 computer – having a fast disk and a few GB of memory.

This forms an upper limit on the processing cost. I suspect that we can do substantially better by crafting the various pieces of software to this specific scenario. The continuum processing is close to CPU bound and may thus benefit from processors with multiple cores. The spectral line processing is I/O bound and would therefore benefit from attention to streaming data.

The next step should be to implement trial versions of these pipelines.

Acknowledgements

I thank Ger van Diepen for helpful discussions.

References

Cornwell, T.J., 2005, ATNF SKA memo series 1,
http://www.atnf.csiro.au/projects/ska/Memoseries.html

Johnston, S., 2006, ATNF SKA memo series 7,
http://www.atnf.csiro.au/projects/ska/Memoseries.html

Staveley-Smith, L., 2006, ATNF SKA memo series 6,
http://www.atnf.csiro.au/projects/ska/Memoseries.html