# Tests on a High Frequency EVLA 100 GByte Simulated Dataset

W. D. Cotton, June 4, 2009

***Abstract*—This memo describes test processing of a high frequency EVLA 100 GByte simulated dataset using Obit on a fast workstation and a node of the casa cluster. Both an 8 GHz bandwidth continuum test at 40 GHz and a 1024 channel line test were performed. Either machine were easily able to process the data in of the order of the observe time. The workstation had substantially better performance than the cluster node.**

***Index Terms*—interferometry, performance**

## I. Introduction

Large datasets will be a prominent feature of the EVLA when it come on line. Some of the impact at low frequencies was explored in Obit Development Memo 8[1] [1]. The expense of dealing with the "w-term" can be large at low frequencies due to the large field of view that needs to be imaged. This memo explores the other extreme, high frequencies where the w-term is far less severe and simple single facet imaging is possible. The following describes use of Obit[2] [2] to process a 100 GByte simulated dataset at 40 GHz and test the feasibility of using a workstation or a single cluster node for such analysis. .

## II. Simulated 100 GByte Dataset

The simulated dataset contains 19 x 10 min scans on "Target" plus calibration scans on "Cal" and "Amp" spread over 9 hours. There are 1024 channels divided into 32 "IFs" equally spaced covering the 8 GHz from 40 to 48 GHz. The Target model was derived from a deep 1.4 GHz B Array image with 131 point sources with the peak 20 cm flux densities and a distribution of locations derived from the 20 cm results but scaled by the ratio of frequencies. There are a total of 2,323,269 x 2 second integrations. 50 mJy Gaussian noise per correlation was added; the model has no frequency dependence. There is about 125 mJy total flux density in the model.

## III. Processing Tests

The tests were of a broadband continuum imaging case using all the data and a "spectral line" case where each channel was individually imaged and then the channel images accumulated into an image cube.

National Radio Astronomy Observatory, 520 Edgemont Rd., Charlottesville, VA, 22903 USA email: bcotton@nrao.edu

Manuscript received ; revised

[1]http://ftp.cv.nrao.edu/NRAO-staff/bcotton/Obit/HundredGB.pdf

[2]http://www.cv.nrao.edu/∼bcotton/Obit.html

### A. Continuum

External calibration was applied and then a single 3600x3600 0.045" pixel facet was formed and deconvolved using a maximum of 5000 iterations of "Cotton-Schwab" CLEAN or a limiting CLEAN flux density of 50 microJy.

### B. Spectral Line

External calibration was applied and then a 3600x3600x1024 spectral cube was formed. Each channel image was deconvolved using a maximum of 2500 iterations of "Cotton-Schwab" CLEAN or a limiting CLEAN flux density of 50 microJy. The processing consisted of:

1) **scatter (SplitCh)**
   Calibrate the data and split into 32 "IF" uv data files.
2) **image (Imager**
   Multiple parallel streams imaging and deconvolving.
3) **gather (MCube)**
   Combine images into single cube.
4) **cleanup**
   Delete temporary files.

With minor modifications, this procedure can spread computing of a spectral cube over nodes of a cluster.

## IV. Timing Tests

The timing tests used mortibus, the Obit development machine, in Charlottesville and a single node (05) of the casa development cluster in Socorro. Mortibus has dual quad core Xenon processors for a total of 8 cores, a clock speed of 3 GHz, 8 GByte memory and a fast disk RAID system. This machine uses the Dell SAS/SATA RAID 5, PERC 6/i Integrated controller (made by LSI) with the Seagate 146GB15K RPM SAS 3Gbps 2.5-in HotPlug Hard Drive. The cluster node has a dual quad processor with 8 2.5 GHz Xenon cores and 8 GByte of memory. The cluster node has two ∼850 GByte hard drives (which James Robnett certifies are working properly).

For the continuum test, a single facet was formed and deconvolved; this took 2 major cycles (the image has a really nice dirty beam).

For the line processing case, two approaches were tested, eight parallel processing streams using a single core each and four streams using two cores each. CLEANing of each line channel took 4 major cycles.

The timing results for the various tests are given in Table I.

TABLE I
TIMINGS

| Process | Machine | Real Time min. | # Streams | Comments |
|---|---|---|---|---|
| Continuum | Mortibus | 81.6 | | Two major cycles, 4581 CCs |
| Continuum | Casa-dev-05 | 124.5 | | |
| SplitCh | Mortibus | 8.9 | 8 | |
| Imager | line | 266 | 8 | Four major cycles |
| MCube | | 31.4 | 8 | |
| cleanup | | 2.2 | 8 | |
| total | | 308 | 8 | |
| SplitCh | Mortibus | 8.9 | 4 | |
| Imager | line | 228 | 4 | Four major cycles |
| MCube | | 31.4 | 4 | |
| cleanup | | 2.2 | 4 | |
| total | | 270 | 4 | |
| SplitCh | Casa-dev-05 | 19.7 | 8 | |
| Imager | line | 366 | 8 | Four major cycles |
| MCube | | 43.9 | 8 | |
| cleanup | | 2.3 | 8 | |
| total | | 432 | 8 | |
| SplitCh | Casa-dev-05 | 19.7 | 4 | |
| Imager | line | 279 | 4 | Four major cycles |
| MCube | | 43.9 | 4 | |
| cleanup | | 2.3 | 4 | |
| total | | 345 | 4 | |

## V. DISCUSSION

The continuum imaging problem was so simple that the bulk of the time was taken in the initial data manipulation, applying calibration and selection, uniform (robust) weighting and conversion to a single Stokes dataset from which to form residuals.

On both machines, the behavior with 8 streams for line imaging was suboptimal, especially so for casa-dev. Apparently with 8 streams, the the whole problem did not mostly fit in memory and causing much I/O. The data for the 4 stream case appears to mostly fit in the disk cache.

On both machines there were files system problems (e.g. files disappearing) for the 8 stream case. This did not happen on mortibus when 4 streams were used but continued sporadically on casa-dev-05. This requires manually restarting Imager to finish the failed runs. On both machines, the data were divided between two disks and there was a notable difference between the run times of streams using each of the two disks. This difference on mortibus was 229 vs. 266 min. and on casa-dev-05 217 vs. 366 min. The run time on the eight stream tests were set by the performance of the slower disk.

In the 4 stream test, the data in each stream were balanced between the two disks to give roughly equal run times for each stream. Note, even though it takes roughly the same elapsed time for a one of eight stream as a one of four stream, the latter does twice as much processing. Neither of the machines appears to deal well with both complex I/O (multiple independent processes) and even modest data volumes. The four way split seems to have worked well and keeps the computing in the CPU dominated regime.

Mortibus appears to have better I/O performance, especially under heavy loading. The scatter and gather operations at the beginning and end show this clearly. The total run time on the cluster node was 50% longer for the continuum test and 28% for the line test whereas the mortibus processors are only 20% faster.

Processing of high frequency data should not require more than a hefty workstation for datasets of this size (typical? for B array). Simple continuum imaging is faster than the observe time and in practice the field of view needed is likely to be smaller than in this test. In reality, wideband continuum imaging, as in this test, will have to account for varying spectral index and perhaps curvature across the field as well as the frequency dependence of the primary beam size. Since the techniques for these are not yet fully worked out, it is not possible to make a timing test but will likely take several times longer than the results presented here. The spectral line reduction took longer that the continuum image but is still of the order of the observe time.

A subsequent test was added in which the continuum imaging also did a single pass of 10 second, phase self calibration using mortibus. This took 3.3 hours rather than the 1.4 hours quoted above for a simple image.

## ACKNOWLEDGMENT

The author thanks James Robnett, Tracy Halstead, Josh Malone and Dan Klopp for their assistance.

## REFERENCES

[1] W. D. Cotton, "Testing Obit with a 100 GByte simulated dataset," *Obit Development Memo Series*, vol. 8, pp. 1–8, 2009.
[2] W. D. Cotton, "Obit: A Development Environment for Astronomical Algorithms," *PASP*, vol. 120, pp. 439–448, 2008.