EVLA Computer Use Plan

Frazer Owen and James Robnett

October 20, 2009

Introduction

March 2010 will be the beginning of EVLA postprocessing when we have turned off the old VLA correlator and will only use WIDAR. At that time most users will be using the OSRO two subband mode which should look very similar to VLA data, only with more channels. If the same integration times are used for OSRO then the maximum data rates are the same for OSRO as for the old VLA correlator. However, more users are likely to use a large number of channels on average so the mean data rate will increase. Nonetheless, all current VLA problems are routinely being reduced outside the AOC, so we don't expect that situation to change until a larger observing bandwidth is allowed for routine observing, supposedly in May 2011.

RSRO observing will be for commissioning the full WIDAR correlator. Initially due to the lack of 3-bit samplers, the full bandwidth at high frequencies will not be available for all antennas. With the recent delay in the 3-bit sampler schedule, we expect to have 8 antennas with 3-bit samplers toward the end of 2010 and all the antennas in late 2011. Thus the data rate at the high frequencies will be constrained by the samplers during almost the entire first phase of RSRO.

During the second EVLA cycle through the configurations starting in May 2011, we expect to increase the routine bandwidth available to users to 2 GHz and to 8 GHz for RSRO But the cycle will start back in the D-array and thus longer integration times will be appropriate. Not until 2012 will we get back to the B and A configurations and shorter integration times.

The S, X and U band receivers will not be completed until the end of construction in October 2012. Thus through the first EVLA configuration cycle, only the low frequencies L and C will be routinely available for most of the array. For the wider bandwidths available for RSRO we expect to work our way incrementally toward full-field continuum imaging not reaching our ultimate goals until early 2013. Most of the lower frequency use will be some version of spectral line imaging, possibly followed by narrow-field imaging once confusing sources are removed. In the second cycle, especially in the B and A configurations starting in February 2012, new imaging algorithms may become available.

The postprocessing problem should not get extremely large until sometime in 2012. Around that time we start to have the potential to do all the sorts of projects the EVLA was designed to do. Before that we will see a steadily increasing workload for our postprocessing hardware but not the extreme problems either at low or high frequencies we have been worried about. However, we must learn how to use our computing resources before getting to the more difficult timeframe when demands by a potentially large number of users will occur.

Experience with the prototyping cluster have led to a definite idea of the architecture we want for AOC computing system which is described below. However, the total processing throughput needed is harder to define. Furthermore we will be limited as to what we can buy by the EVLA construction budget and the longer we wait for a purchase the more we will get for our limited dollars. Recently experiments with single nodes working on large datasets plus experiments doing data compression on-the-fly for imaging have suggested that larger problems are feasible on individual workstations. Furthermore, we will continue to have access to significant time on the existing prototyping cluster, although its first priority will continue to be testing. It seems likely that we still will need a new production cluster but not quite as early as we first thought, perhaps Q4 of 2010.

We are proposing a staged, learning process to make the wisest use of resources. We will begin in March 2010 by relying on individual workstations and the available time on the existing prototyping cluster for RSRO. In summer 2010 we will upgrade the system including a purchase of a first production cluster. Based on experience with that system during RSRO we will buy a final EVLA construction cluster in late 2011 which will be well-matched to our needs for the full EVLA.

What follows describes a basic model for the postprocessing. A basic model for the hardware is described. Some possible modes of its operation are outlined and some remaining decisions detailed. A list of development tasks needed for the user model and a first order schedule for the deployment of the system is given.

The Basic Model

The basic hardware model for EVLA postprocessing at the AOC (as shown in figure 1) consists of

1. Central NGAS data archive (currently exists)
2. Central shared disk array
3. Some number of high end workstations (or single nodes on the cluster)
4. A cluster computer
5. A fast network connecting all of the above

Access to data will involve a number of potentially redundant reads and writes as data moves from the archive to one or more workstations and the post processing cluster. A significant issue is the total time spent simply moving data sets from system to system. We estimate needing a transfer rate of 600 GB per hour (150MB/s) out of the archive and roughly half that rate onto a cluster node or workstation to avoid load times dominating total data reduction times.

To that end we plan on creating a parallel file-system based on LUSTRE (figure 1) that can be uniformly accessed from the NGAS based archive, the post processing cluster or individual workstations. Such a file-system will eliminate the need for redundant local writes of UV data sets. Further study as to performance and scaling issues are needed to properly design such an array to achieve the target data rates. In addition, portions of the AOC network will need to be upgraded to support a new dedicated network for the disk array, cluster and workstation data traffic.

The user at the AOC would be assigned a workstation or have one in his or her office attached to the high speed network. Basic external calibration, editing, examination of images and other interactive tasks would be performed on these workstations. In some cases the entire reduction might occur on the workstation. Trial reductions of subsets of the data would be performed to verify the setup before submitting a job to the cluster. Since the LUSTRE storage array is uniformly accessible from typical workstations and cluster nodes it's possible a workstation could in fact be a single node on the cluster. The life span of data on the array is not a function of how it is accessed.

In the event further reduction on the cluster is needed, a job would be set up on a workstation and passed to the cluster. Supporting data such as calibration and editing tables and any other information needed, e.g clean box tables, would be written to the storage array from the workstation and be accessible by the cluster. The job would be queued to run on the cluster using some priority scheme. When the cluster job finished the output images, tables, etc would be written back to the storage array and thus made available to the user's workstation for examination and any action necessary including more use of the cluster (some thought will need to be given to appropriate permissions and access). The user would then examine the results perform any smaller tasks on the workstation and, if necessary, submit another

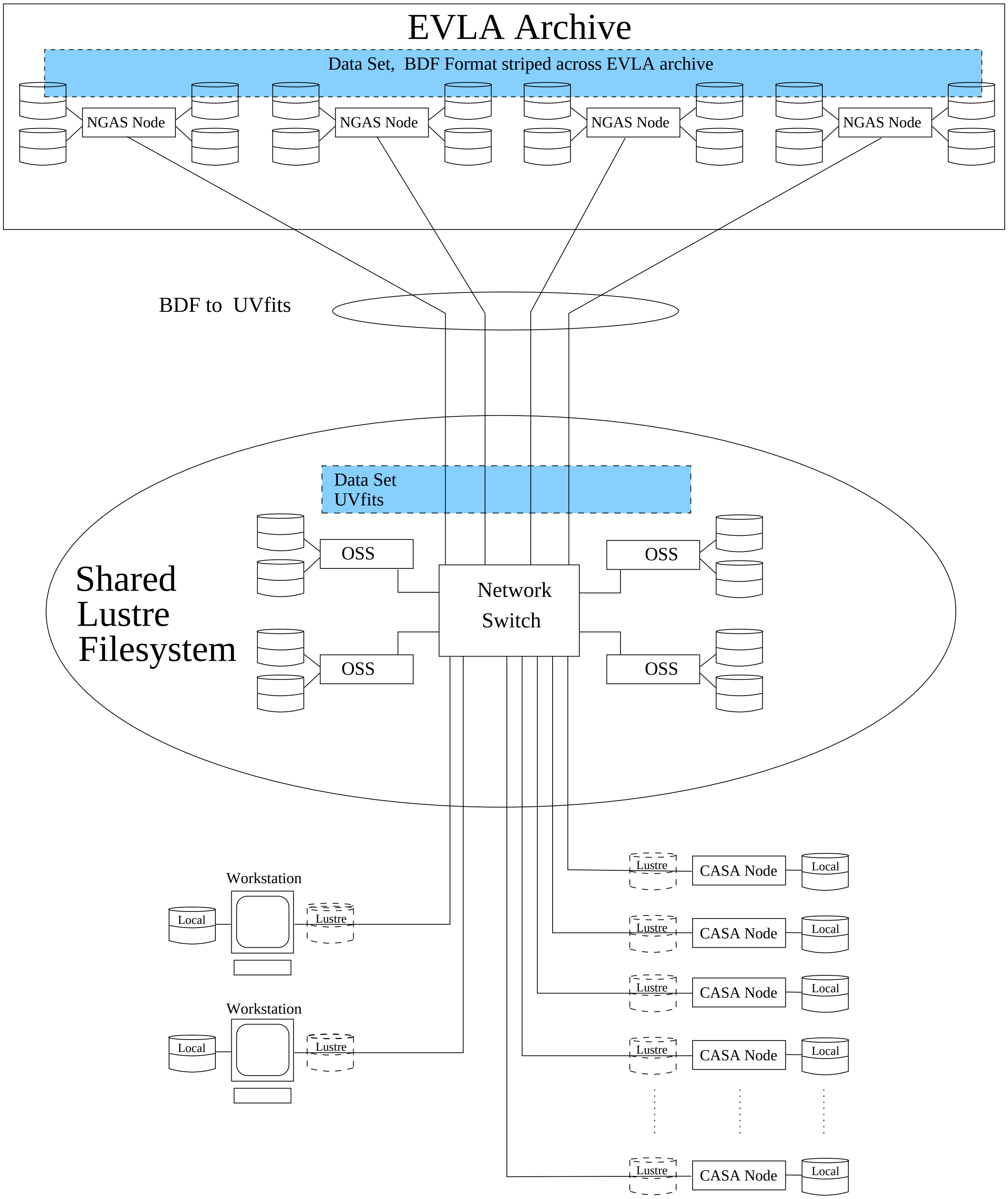

Figure 1 Schematic diagram showing AOC computing layout

job to the cluster to continue the reduction. The number of workstation-cluster iterations a typical user would go through will be a driving factor in sizing for the optimum cluster.

We will specify a first cluster by the end of November 2009, but continue to refine our specifications as long as possible in order to get the best, first production cluster. We now think the purchase can wait until summer 2010. Even this cluster should be looked upon as a learning system from which we should be able to gain the necessary experience with the user community in order to purchase a "final" cluster, well-tuned to the user community at the very end of the construction project. A more detailed specification of infrastructure, hardware and network requirements can be found in James Robnett's 5 year Infrastructure Plan (under development as of 10/20/09).

Users not at the AOC

Most users will likely prefer to work from home. However, many users may not have the computer facilities to complete all the steps in the reduction with their home computing resources. Two obvious models exist for dealing with this situation:

1. The user does all their reductions on NRAO computers by logging in remotely and being assigned an workstation/node. In this case, no data transfers are necessary.

2. The user does as much as possible on their home computer using our cluster only when necessary. In this model NRAO transfers a copy of the uv data for a project to the user. The user edits, does any external calibration needed on the dataset at home and submits any very big jobs to the AOC cluster over the Internet. The big job submission would include the necessary tables or images from the user's computer but not the uv data. The uvdata would be read from the archive and combined with the input tables and images. When the cluster job is complete, the user would collect the output images and tables and continue the analysis on his or her home computer. Over time such a system would probably evolve into more self-sufficient users with clusters and home (or on their desks) and less use of NRAO computers. Perhaps as an intermediate step in the evolution this system could be broadened to regional computer centers to which NRAO would send the piece of the uvdata archive that was needed for projects intended to be run at those sites.

Option 1 is simpler. However, it puts a much larger management load on NRAO, limits the flexibility of the reduction process for the user because of needing to schedule time on NRAO's computers and would require NRAO to buy more computing hardware. It seems likely that most users would prefer to do as much as possible at home on their schedule. However, we may want to have option 1 for some users.

Option 2 might also require more management than we would like if the cluster is oversubscribed. We would need some way to establish a priority system for jobs submitted to the cluster, rather than first come, first served. In either case we are likely to be limited by the budget available for computing. Option 2 would put more of the load on the community and corresponding less on NRAO. This choice is an important management decision which we do not propose to make in this memo.

What we need besides the hardware

Besides the computing hardware, we need a plan to provide the framework for the use of the cluster in such a model. This plan needs to be developed by a working group which should be formed by management from appropriate computer operations and software development staff. We need:

1. A method of transfer of the uvdata to the user, either locally in the AOC or remotely. It seems likely that this will require some hardcopy of the uvdata to be shipped to the user or as an Internet transfer when that is practical. We need a detailed plan.

2. A queuing system for jobs sent to the cluster.
3. A plan for how to pass the information in tables from the user to the cluster, or perhaps the shared disk array, in an efficient way.
4. Integration of the entire system.

Schedule (dates can be modified)

Below is a notional plan for evolving the use of the cluster:

November 2009: Produce full, initial specification for a cluster, the shared disk array and the post processing network we could buy immediately based on what we know then.

March 2010: Start with existing AOC individual workstations, both public and in scientist offices and the existing prototyping cluster.

May 2010: Upgrade workstations to handle larger datasets.

July 2010: Purchase initial production cluster, upgrade network, disk array.

September 2010: Start regular queue use of cluster in AOC

May 2011: Begin experiments with outside users using the AOC cluster.

October 2011: Purchase second generation cluster for full EVLA capabilities.

January 2012: Begin open use of cluster by outside users.

Note that in March 2010 we will have all the elements of the final system described in figure 1 in place. The proto-typing cluster will be available for some processing, although its first priority will continue to be testing. The July 2010 date for the purchase of the first production cluster is driven by our perceived need to have the cluster ready for the B configuration which begins in mid-October 2010. The October 2011 date for the final cluster is driven by wanting to be ready for the B array which begins in February 2012.

Learning from the first production cluster

We have learned a lot from the existing proto-typing cluster about how clusters perform with simulated data and how we need to develop in our software to take advantage of this type of resource. However, we cannot simulate how the user will interact with real EVLA data being reduced on the type system we are planning. This learning process needs to happen during RSRO and thus the first production cluster is aimed not just at reducing EVLA data but allowing us to form a better picture of what we need in the long-term.

The break point between experiments entirely appropriate for a single workstation and those which need the cluster is hard to specify quantitatively. Which stages of the reduction of a big problem are best done on the cluster is unclear in detail. For example, some automatic flagging and/or interference removal could require the cluster but we hope the routine case could handled by a workstation. When the cluster is required, for example for imaging and self-calibration using different parameters, the typical number of iterations on a single dataset using the cluster is hard to predict.

Probably the most important unknown is how users will prefer to work when confronted with a large EVLA dataset. Many may prefer to do as much as possible on a single workstation they control. Others may want to use the cluster as much as possible and/or have NRAO staff do most of the work. Some practical equilibrium is likely to develop which is hard to predict.

Given the uncertainties, especially the human ones, we need to look at the first production cluster as a learning experience. Based on the experience we gain before we purchase the second cluster, nominally in October 2011, we should determine what we really need for the long-term for our users. We don't think we can specify this now but with 16 months of experience with RSRO we think we will be able to made a wise decision.

User Support

As with the computer hardware, it is hard to see clearly what will be needed for user support, especially in the early years. Perhaps user support will be easy and through scripts and a simple extrapolation from the VLA all our users will be able to continue working with EVLA data as they have been doing for a long time. However, one lesson from EVLA so far is that generally things are harder than anticipated especially when they involving computing. Thus it seems prudent to plan for a case that is not just ”blue sky”. What follows could be too pessimistic but it that turns out to be the case then I don't think we will waste the resources proposed. We need a clear plan in user support.

It would be nice if all the EVLA processing was transparent to the user so that they could continue to do things the way they have in the past. At least the larger databases force some new wrinkles on them. However it seems likely that there will be more of a learning curve than that. In the 70's and 80's everyone expected to come to the VLA site at least for their reductions. By now people have gotten out of that habit. Nonetheless we must face the likelihood that all of our users will need to be retrained, including the local staff. One way or another this will take a lot of AOC scientist time. It seems likely that the only effective way to accomplish this task will be hand-holding. That is, like in the original VLA days, a staff member will need to guide the outside user through a learning project.

Ideally we would write a lot of clear documentation which would allow the outside user to reduce EVLA data using a cookbook approach. We certainly need to write such documentation. However, the time in between the learning process for our staff and the outside user may not be very long. The most efficient way to deal with this situation would be to have each new user (or at least one person from each group) to come to Socorro to be led through a first project. If we need to do the coaching remotely, we are going to have our coaches on the telephone or email a great deal of the time. For very different projects, this process might have to happen more than once (e.g. wideband continuum and high dynamic range line work). One way or another we are going to have to have a moderate size group of local experts covering each different class of experiment as hand-holders or as documenters or both. We don't need very many such people now since we have a large community which knows how to use the VLA and run AIPS.

An estimate based on past experience when the VLA was young is that we need a core of 5-6 such astronomer helpers in 2011-2013. That group could be supplemented with another 5-6 part-time, who are less expert but could fill in when needed. It seems likely we will ultimately need to train 100-200 users. Typically we might expect ~ 10 new outside users per month in this model. We also can anticipate a lot of follow-up help to these same users when they try to reduce data at home. It seems difficult to avoid this situation given our need to bring up the EVLA in a few years timescale.

Alternatively we could reduce all the data ourselves and send users the finished images. We probably will not have the resources to do this and it would be better to develop a set of experts. We may also face the problem that the community may be unwilling to put in the effort to learn. Based on past experience this will not be a problem with our core users. It seems unlikely that we can know how to handle the user training problem until we get to it. So we should make every effort to have adequate scientific staffing for training. This problem could be the ultimate EVLA Achilles heel.