NAASC Computing Strategy

NAASC Memo 101

Author: Mark Lacy

Date: 9th August 2010; revised 15th August 2011

ABSTRACT

This document describes the computing strategy of the NAASC from ALMA commissioning/science verification through to nominal operations.

1. Background

As ALMA ramps up, computing requirements will accelerate in step with the continually increasing sizes of ALMA datasets. The pipeline cluster will be procured for the Santiago Central Office (SCO) on 1st January 2012. Until that time, the only dedicated resources for ALMA data analysis in Chile will be desktop machines, running scripts rather than an automated pipeline. At the NAASC, we are therefore preparing to support Early Science (ES) operations by ensuring we have sufficient computing power to deal with the likely processing and storage requirements through ES and into full operations. This document summarizes the data challenges ahead and how we plan to meet them during the phases of ALMA ramp up.

The ramp-up outlined below is subject to revision if, for example, the data volume turns out to be much lower or higher than predicted. In particular, as the computing clusters and NGAS storage can be built up over time, they can respond flexibly to the actual data load.

We have divided the document into three main sections corresponding to Commissioning/Science Verification (CSV), Early Science (ES) and Operations. Note that hardware orders typically lead the start of each of these phases by a few months to allow for hardware purchase and setup.

2. Commissioning and Science Verification (August 2010 - August 2011)

2.1 CSV Data rate

Starting early in 2011, we began gaining experience with a small number of Commissioning (C) datasets, followed by a more substantial effort to understand and document Science Verification (SV) data. Data rates were low enough that transfer via internet or small capacity hard media sufficed. The Observatory would like to make the SV datasets publicly available using the ALMA Science Archive (ASA), but the ASA at the ARCs did not have the capability to do this in time for the first CSV datasets, so data was transferred by hand to the NAASC.

2.2 CSV Archiving and storage requirements

Each CSV dataset is ~10-100GB in size, assuming 10 datasets results in <1TB of data. This small amount of data was easy to manage. Our strategy was that a server be provisioned where up to ~1TB of raw data could be staged, with a further few TB of space for storing intermediate analysis products. NAASC scientists analysed analyze the CSV data on their desktop machines (or later on the 8-node cluster), and stored the results on the server as required.

To aid in the dissemination of the ES data, the NAASC computing group produced a torrent download engine, in addition to allowing regular FTP.

2.3 CSV Data processing requirements

CSV datasets are capable of being processed on desktop machines or single cluster nodes. We will use our experience with processing and benchmarking the CSV and initial Early Science data to derive recommended specifications for desktop machines that are capable of analyzing ALMA datasets, which users can buy for their own processing needs. For this purpose, we will provision a suite of desktop computers, ranging from relatively affordable "low-end" systems that should fit the budget of most university investigators, to more capable (and more costly) "high-end" systems (Appendix 1). Both Linux and MacOS systems will be purchased and benchmarked, with the results posted so that the community can make reasoned decisions based on their science goals. When this benchmarking process is complete these machines will be used for visitor support.

2.4 CSV Shopping list

- One file server with ~5TB of storage for serving CSV data and storing analysis within the NAASC. Requisitioned September 2010.
- 6 multi-core desktop machines with >8GB memory per machine and >1TB of hard disk, range of processors, operating systems and memory to help determine recommendations for users' own purchases. Purchased November 2010-September 2011. (See Appendix 1 for details.)

3. Early Science observations (August 2011-Sept 2012)

3.1 ES Data rate

Early science data will be taken starting when there are 16 available antennas and will continue until ALMA inauguration in September 2012, by which time most of the antennas should be installed. Data rates will ramp up approximately as the square of the number of antennas (though only about half the available time will be available for ES observations due to ongoing commissioning tasks). Data rates will also be affected by available correlator modes and capabilities. For example, integration times may not be optimized, and it will not be possible early on to select multi-resolution modes. The fact that no data will be taken in wide array configurations will, however, allow long integration times, and hence reduction in volume (if this can be implemented before the data are written to the archive). Data rates are thus highly uncertain, but at the start of Early Science are likely to be around 20TB/yr, ramping up to 200TB/yr by September 2012 (Appendix 2).

3.2 ES Archiving and storage requirements

The total data volume during ES is similarly difficult to estimate due to the uncertainty in the antenna delivery schedules, but a reasonable guess for the whole

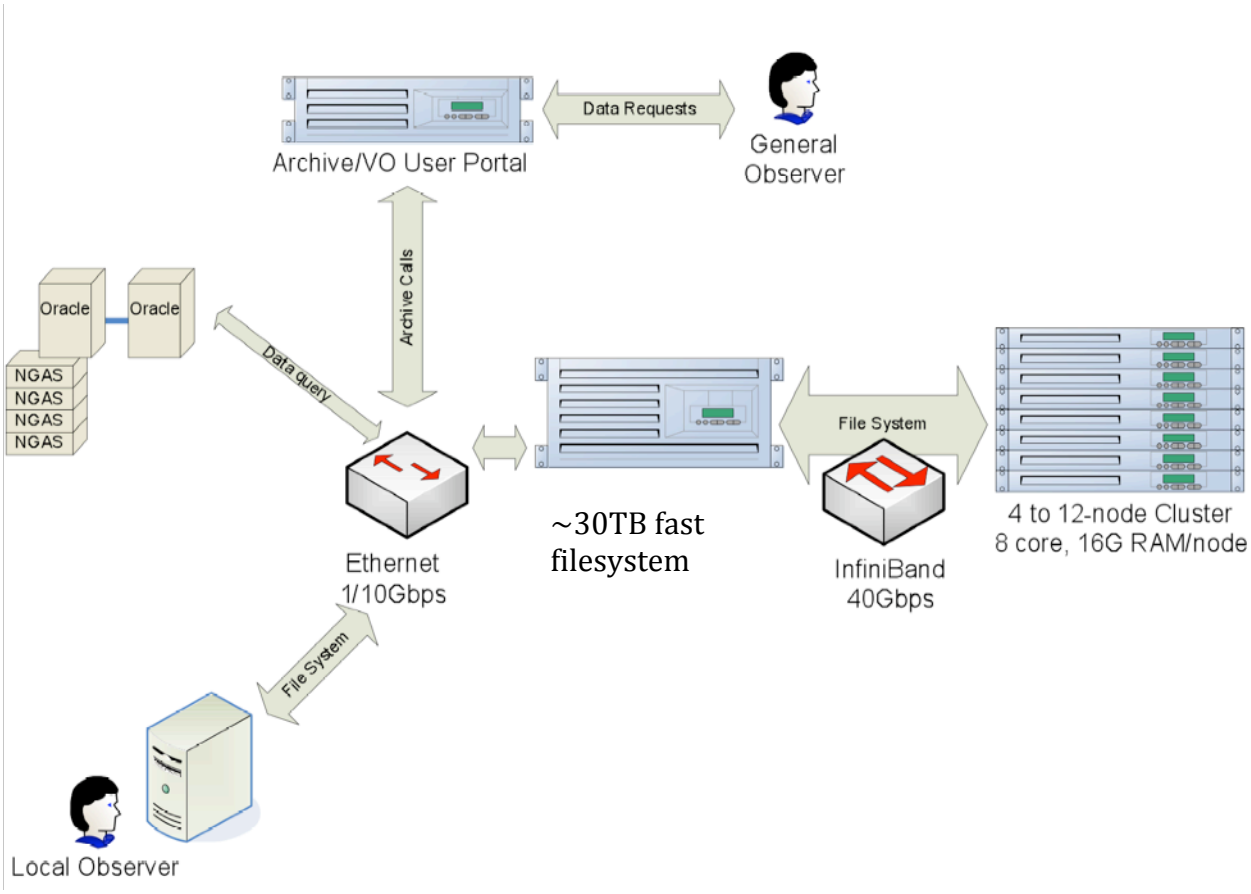

Figure 1. Early Science cluster configuration

15 months of ES would be 100TB. It is planned that, by the time of ES, the archive database replication will be working, and data mirroring enabled. At least a crude archive interface will be available to users to download the data. Additional, non-archive user storage will be needed to support visitor and staff needs.

ALMA data processing requires a large amount of temporary workspace. Data processing begins with a doubling of the data volume when the 32bit ALMA Science Data Model (ASDM) data is converted into the 64 bit Measurement Set (MS) used by CASA, and many temporary tables and ancillary images are produced. With this in mind, we have purchased 60TB of storage for use as fast scratch storage by visiting and internal NRAO ALMA users while analyzing ES data. Two further Lustre nodes will be added in Fall of 2011 to increase this storage to >100TB.

3.3 ES Data processing requirements

Early Science will present some of the first truly challenging datasets. The largest individual datasets are likely to exceed 100GB, and will require computing power beyond that of a typical desktop computer. EVLA memo 132 discusses the processing requirements for a 100GB dataset, showing that, on a single CPU, the data take 30 times longer to reduce than to observe. Also significant is the fact that the breakdown of I/O:computing time was 40:60, implying that a fast I/O is almost equally as important as CPU power. (Although this memo is now somewhat out of date, as speeds have increased, the ratio of I/O to computing time is roughly the same.)

Efforts to parallelize the most computationally intensive parts of the CASA code are underway, and the first parallel code, parallelizing the simplest modes of the clean algorithm, should be available in the Fall of Calendar Year (CY) 2011. Assuming this is effective, in the sense that the processing time scales inversely as the number of cores, current tests suggest that a single, modern, multi-core machine should be capable of processing data at a rate sufficient to match the average rate at which data is taken during ES (though this needs to be confirmed). Thus single desktops or servers may be an acceptable solution. However, they may not be adequate to deal with the largest datasets from ES, or be efficient if multiple iterations of processing are required.

At the NAASC, we have therefore opted to build a small cluster during ES (ordering the first 8 nodes in the third quarter of FY2011, and building up through CY2011), with the idea that it can be used as a basis for the full pipeline cluster, as well as being capable of reducing ES data in less than the time that it takes to observe (Figure 1). The specifications will be determined from the first 1-2 months of data processing experience with the EVLA cluster currently being built in Socorro, and we will use their experience to determine the exact setup (file system, interconnect, RAM per node etc). NAASC personnel have visited Socorro to help with the setup of the EVLA cluster, and learn directly from their experience.

We expect the cluster to be able to keep up well with the NA component of the data from ALMA, and we can add cluster and file system nodes as the data rate builds up during ES. Visitors with ES data would be able to access this machine in Charlottesville to reduce their data, or would use the desktop evaluation machines described in Section 2.

3.4 ES Shopping list

- 60TB of fast centralized storage on two Lustre nodes for local users scratch/workspace, with a further two nodes to be added during ES.
- 8 cluster nodes for a prototype cluster, with a further 48 added as needed during ES and interconnect with Lustre nodes (specifications informed by work on the EVLA cluster).
- Cluster management system (includes scheduling functionality).

- Additional NGAS nodes as required.

The cluster will be built out from an initial eight node system, purchased during the Q3 FY2011. Additional nodes will be added through CY2011, with sufficient lead time to allow time for set up and tuning before the first large quantities of ES data arrives. Purchase could be delayed if ES is delayed significantly. The centralized user storage will also be built up over time, beginning in Q3 of FY2011.

The pipeline is not operational during ES, we anticipate that most data (re)processing in Charlottesville will be carried out using remote VNC desktops on individual cluster machines. Fast infiniband connection to the Lustre file system, and the striping of the data across multiple disks in Lustre makes the cluster nodes significantly faster than even a high-end desktop. VNC is much more efficient than an ssh connection as e.g. individual plot data points do not need to be transferred over the internal network. Appendix 3 discusses VNC technologies for remote desktop sessions.

4. Full array operations and the pipeline cluster (Oct 2012 onwards)

4.1 Operations data rates

The ALMA Operations Plan, version D (AOPvD), specifies 6MB/s (200TB/yr) as the mean visibility data rate that will be archived by ALMA, based on calculations made from the Design Reference Science Plan (DRSP). (The mean volume of the image cubes produced by the pipeline and archived will be a small fraction of that (<10%).) The maximum data rate, limited by the data control interface boards, is 64MB/s, and the maximum data rate that can be transferred over the proposed fast internet link between the OSF and Santiago is expected to be about 12MB/s. Thus the actual data rate during operations may be significantly higher than 200TB/yr. We have therefore assumed an aggregate data rate (including imagecubes) of 500TB/yr.

4.2 Operations processing requirements

In its final configuration, a typical 12hr ALMA dataset will be several hundred GB, and a large dataset 2TB or more (EVLA Memo 132). The CIPT is charged with procuring a production pipeline analysis cluster for the SCO in January 2012 based on specifications developed during 2011, along with the pipeline software, with the aim of having a working pipeline one year after the start of ES (August 2012 based on the current schedule). Nearly all ALMA data will be processed through the pipeline at the SCO (some large datasets may require special treatment beyond that feasible with the SCO pipeline). It is likely that the data volume from ALMA means that most datasets will only get a single pass through the SCO pipeline.

The cluster purchased by the SCO will be based on the EVLA experience, and those of the ARCs with ALMA Early Science data. At the NAASC, our plan is to build up our ES cluster to achieve a comparable performance to that at the SCO (Figure 2). The functions of the pipeline machines at the ARCs are to help develop pipeline

heuristics, rerun problematic datasets through the pipeline, and perform any bulk reprocessings required by the invention of new algorithms, or correction of pipeline errors. In addition, the NAASC, as part of its full science function, will use the cluster perform user-driven reprocessing of ALMA datasets. Incremental reingestion of post-processed data into the ALMA archive will be subject to international agreement, but is likely to occur for project-sanctioned reprocessings (ALMA Science Operations: DSO and ARC coordinated acitivities vers. A2, 9/9/2009). Depending on the prioritized load from reprocessings, at the NAASC we will allow visitors some access to the central cluster for the purposes of processing their data, though whether this would be solely via scripts, or whether we would allow interactive use is still to be decided. Ultimately we envision users being allowed to process their data remotely, either via supplying pipeline parameters through a web interface, or by using a secure VNC connection. We will use our local experience with EVLA processing to determine which of these two solutions is best.

The specification for these clusters is thus unknown at this time, but likely to have around 64 multi-processor nodes, with a large amount of RAM (~24GB) per node. Assuming the processing problem scales linearly with the number of nodes, such a machine should be able to process ALMA data as fast as it is obtained.

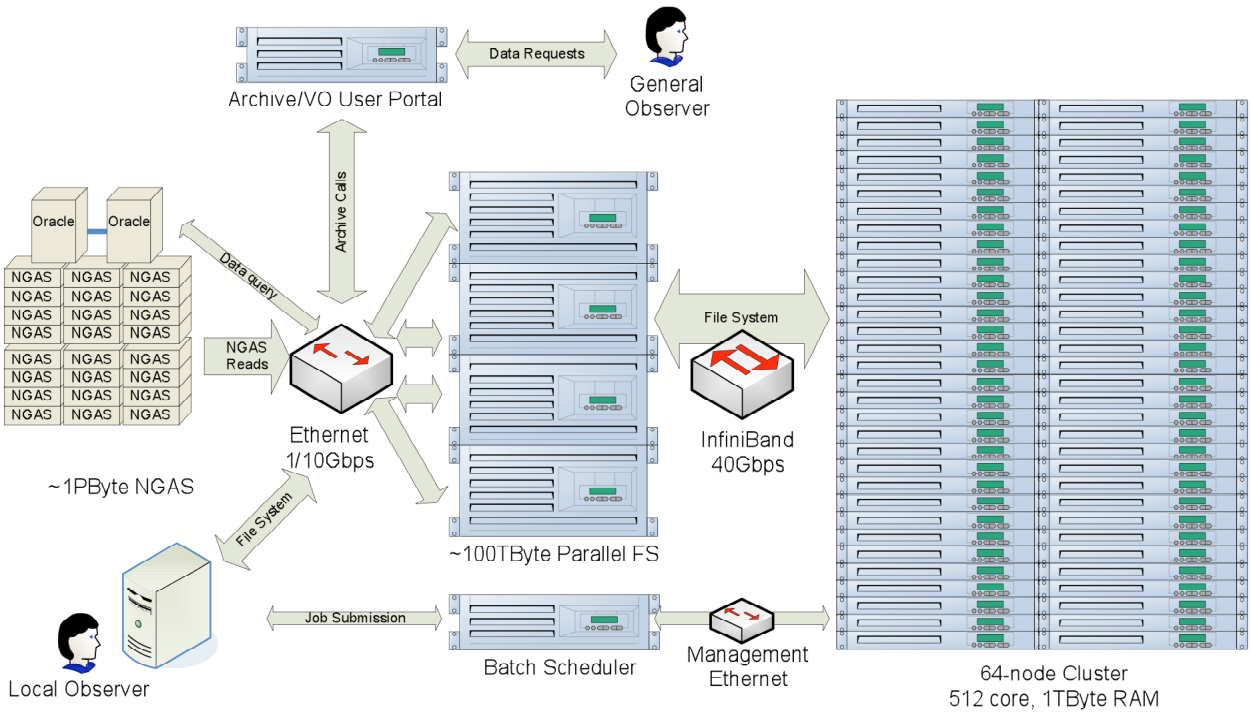

Figure 2. Operational cluster configuration

4.3 Operations archiving and storage requirements

We have assumed a growth rate of up to 500TB/yr for the NAASC archive during operations (compared to the estimate in the AOPvD of 200TB/yr). This could be

higher if, for example, we decide to supply measurement sets at the NAASC by performing the ASDM to MS conversion ourselves and storing the results. Fortunately, the NGAS system is scalable and even the total of ~1500TB/yr needed if we perform this conversion on the full 500TB/yr of ASDM data and store the results should be within the capabilities of the NGAS system.

We will also add a further 70TB of non-archived scratch storage for visitors and internal users for a total of 100TB.

4.4 Operations shopping list

- Full pipeline cluster and interconnect as specified by CIPT (upgrade from ES cluster to full cluster, assuming specifications are consistent). Order early 2012 to be up and running by the summer. Also may grow with time through 2013 as antennas continue to be added.
- Additional NGAS nodes as required.

5. Risk and mitigation

Should our cluster prove insufficient or unable to process all the data from full operations, we will negotiate for use of National (TeraGrid/XD) supercomputer resources, and/or resources from our collaboration with UVa for the larger datasets, and rely on our cluster and/or further powerful desktop machines to process the smaller datasets. Decision point: June 2012.

Appendix 1. Costing of example desktop/single machine solutions

1) "Mid-range" Mac: Apple Mac Pro Quad Core; 2.66GHz processor, 8GB RAM, 4x1TB disks, 24" flat panel display, $4550
2) "High end" Mac: Apple Mac Pro 12 core; 2x2.66GHz Six core processors, 24GB RAM, 4x2TB disks, 27" flat panel display, ATI graphics card $7100 (less educational discount)
3) "Low-end" Linux desktop: Dell T3500; Dual core 2.8GHz processor, 24GB RAM, 2x1TB drive, 19" display: ~$3000
4) "Mid-range" Linux desktop: Dell T7500; Dual Quad core 2.26GHz processor, 24GB RAM, 3x2TB drives, 2x19" monitor: ~$5000

Note that we have no "high-end" linux recommendation pending further knowledge of CASA parallel performance.

Appendix 2. Data volume calculation methodology

Current assumptions via B. Glendenning:

In full science -

"Typical" 12hr ALMA dataset is 250GB (6MB/s for 12hr)

Peak ALMA dataset is 2.5TB (60MB/s for 12hr)

Scaling to early science, assuming 16 antennas (occupying one quadrant of the correlator) and typical ES correlator modes results in a 10x smaller ([50/16]^2 ~10) dataset than the full 50 antenna array. So we assume, both for ES and late SV that an ALMA dataset would have about 1/10 the volume of the full array. However, this does not allow for the fact that the data output settings (averaging time, spectral windowing etc) may not be optimized.

Appendix 3. VNC technologies

VNC servers and clients are commonly installed on Linux systems, and allow remote access to an external machine's desktop. Multiple external sessions can be run.

For Macs, a free VNC client is available ("Chicken of the VNC"), making it easy to use a remote desktop running on a Linux host machine. Unfortunately MacOSX does not allow multiple remote desktops, so it is not possible to run independent sessions off a Mac via VNC, but mirroring of a remote desktop is allowed, and fast-switching of user accounts can be used to mimic a remote desktop login. We do not recommend a Mac system if there is a requirement to run multiple remote desktop sessions on it.

Appendix 4. Initial 8-node cluster specs. From J. Robnett (used by JAO and NAASC)

Compute nodes ($26k):

8xDell 410 model 1U servers with dual 2.4GHz hex-core E5645 Nehalem processors.

24GB of memory (2x3x4GB)

Local OS disk only

40Gb QDR Mellanox NT26428 PCI-E 8xHCAs for fast access to Lustre storage.

Lustre filesystem ($25k):

1xDell 410 1U server with two 250GB internal disks (RAID-0 mirror of OS and metadata target), 8GB memory and a 40Gb QDR Mellanox NT26428 PCI-E 8xHCA to act as the Metadata server.

2x4U 24 disk chasses with Superlogics X8DTH-I motherboard, 7 PCI-E 8x slots, dual E5520 Xeon processors, 4GB memory, 40Gb QDR Mellanox NT26428 PCI-E 8xHCAs and redundant hot swappable 1200W power supplies to act as Object Storage Servers

8xObject Storage Targets (RAID arrays) with an eight port 3ware 9650SE raid controller attached to one of the PCI-E 8x slots and 6 Western Digital WD2003FYYS 2TB hard drives.

Infiniband network ($10k)

36 port Mellanox IS5030 QDR switch plus a Fabric Subnet management license, the

individual HCAs on each node and twinax cables of various lengths to connect each node to the switch. Cables part numbers are MCC4Q26C-00X where X is the length in meters; 4 meters are used as an example. (Note 15m max cable length for infiniband)

Total cost ~$61k, excluding infrastructure items such as racks, power, PDUs, AC etc.