NRAO Whitepaper Submission to the NSF Large Facilities Cyberinfrastructure Workshop

B.E. Glendenning, Assistant Director, NRAO Data Management and Software Department

1 NRAO TELESCOPES

The National Radio Astronomy Observatory (NRAO, https://public.nrao.edu/) operates the Karl G. Jansky Very Large Array (VLA) near Socorro New Mexico, and is the operating partner (Executive) for the North American part of the Atacama Large Millimeter/Submillimeter Array (ALMA), which operates at a high site near San Pedro, Chile.

Both telescopes are very general purpose. Telescope time is allocated based on a peer-review process from many sub-fields of astronomy. Hundreds of PI groups per year get data, and in addition once the proprietary period has expired (usually one year), the data may be used by other groups for Archival research.

Both telescopes are radio interferometers, which operate by coherently combining the signals of the relocatable antennas (27 for the VLA, 66 for ALMA) in complex central electronics (notably the correlators, which are approximately 0.1 Exa-Op very parallel special purpose supercomputers) which produces raw data, essentially a noisy (electronics, radio-frequency-interference, atmospheric and other environmental effects) irregularly sampled spatial Fourier transform of sky "stacked" over separate frequency channels for up to 4 polarizations.

The electronics are capable of sustaining 1 (VLA) and 16 (ALMA) Gigabytes per second of raw data output, although the data rates are usually averaged down (in time, and frequency) to a small fraction of that (typically 25 Megabytes/second for the VLA, and 6 MB/s for ALMA). This averaging is done both to reduce the computing that is needed, and because many times the science application do not need high data rates. However there are some classes of science observations that are not made because computing capacity is not available.

The raw data is turned into regularly gridded 2-4 dimensional images (axes: position on the sky, frequency or Doppler velocity, polarization) using multi-million line of code software systems produced by the NRAO and our partners. These images (currently: Giga-pixel, coming Tera-pixel, Possible: Peta-pixel) are then typically processed through analysis codes (both produced by NRAO and the wider community) to enable the science to be extracted from the data.

2 CURRENT NRAO COMPUTING PARADIGM

The raw science data from each telescope is buffered at the telescope site (to allow for network outages and periods of high data rate observing), from which it is transferred and ingested into the master archive (in Santiago in the case of ALMA, Socorro NM in the case of the VLA). In the case of ALMA the data is then replicated from the master archive to the "regional" archives, which for North America resides at Charlottesville Virginia. Through an archive search web interface the raw data may be

downloaded by operations staff and the PI group that proposed the observations (after QA in the case of ALMA). The raw data may be freely downloaded by anyone after the (typically) 1-year proprietary period has expired.

After the raw-data for the entire project has arrived in the archive (this could take several different observing sessions), "pipelines" are executed which automatically make derived data products, currently flagged and calibrated raw data for both telescopes, and reference images for the case of ALMA. After some QA is performed, these data products may be downloaded by the PI groups, or by anyone after the proprietary period has expired. NRAO has initiated a "Science Ready Data Products" (SRDP) project to improve the quality of the automatically generated data products, with a goals that: the images should be directly usable for science, to improve the user interfaces, and to allow a human to be in the loop to optimize via high-level guidance the derived data products to be well suited for use in answering particular science questions.

At the moment, almost all VLA derived data products, and many ALMA ones, which are used for the actual science analysis are produced through the manual (including ad-hoc Python scripting) execution of programs from suites of data processing, analysis, and visualization tasks produced by the NRAO. These programs are developed by the NRAO with significant contributions from our ALMA partners, and total about 3M SLOC. This software is available under an open source license, although the NRAO generates executables for common Linux variants and recent versions of MacOS.

The software is executed at a combination of NRAO and user facilities. Our software is downloaded several thousand times per year for use by users (laptops through small clusters). In addition the NRAO allows our users to use our in-house computing facilities through a reservation system. Although our resources are relatively modest (150 16-core compute nodes, 2 PB of fast Lustre filesystem with Inifiniband interconnects), they are well tuned to our software stack, have fast access to the raw data archives, and we allow them to be used interactively (we also have batch queues). That is, they are convenient to use and very suitable for modest problem sizes. Our computing resources are used by a few hundred PI groups per year.

We have experimented with commercial cloud providers (AWS) and national supercomputing centers (XSEDE), but have not made extensive use of either yet, nor have our users.

Key CI improvements areas we would identify are:

- In-the-cloud Elastic, Interoperable, Data Center accessibility
- Machine learning applications (vs. ad-hoc expert knowledge capture in scripts)
- Software sustainability infrastructure
- Visualization and information extraction from multi-peta-pixel multi-dimensional image data

We look forward for the opportunity to discuss these topics at the workshop.