Scaling NRAO to the Future

Brian Glendenning

NRAO Data Management and Software Department

Current NRAO Computing Model

- Raw telescope data rates throttled to average of 10-25 MB/s (capable of 1-16 GB/s) to enable computing on in-house and user facilities
- Data buffered at the telescopes, and then ingested into replicated, permanent, archive at NRAO locations
- Mix of manual and automatic data processing to turn raw data into science images
  - 3M SLOC software packages (25+ years old)
- Computing at NRAO (3k core clusters, Lustre filesystems with Infiniband interconnects)
  - Used for NRAO processing and >200 user groups per year
- Computing at user institutions
  - Enabled by NRAO software (thousands of downloads per year)

Opportunity – Data Center in the Cloud

- Increase capacity: lift data rate limits for telescope observing
- Remove IT infrastructure burden on NRAO and (especially) end users
- Cater to bursty processing, e.g. thousands of hours for VLA (all) sky survey (VLASS) in one season, reprocessing when algorithmic improvements available
- Allow for more expensive algorithms (needed for VLASS already)
- Issues
  - International standards interoperability
  - Durable agreements (we have all our data back to the late 1970s)
  - Relatively high storage / CPU provisioning required
- Opportunity: HTCondor technology, OSG/CHTC/Consortium collaboration

Opportunity – Machine Learning

- Currently encode "expert knowledge" in automated data reduction processes via Python scripting using straightforward logic based on computed values from raw and processed data
- Seems like a ripe time for exploring machine learning techniques for:
  - Flagging (bad data excision), calibration, image production
    - (Some modest/promising experiments in flagging radio frequency interference)
  - QA assessment of raw and processed data
  - Data mining from our accumulated (40 year) archive
  - Pattern recognition / source finding in 2+ dimensional data (particularly for low signal/noise sources)
  - Fault detection/isolation in our telescope hardware/online software

Opportunity –Sustainable Software Infrastructure

- Our software has been developed over decades
  - Our "new" software is 25 years old, our "old" software is nearly 40
- Software is our capital expense
- Hard for us to tune our software when the underlying platform/paradigm changes every few years
- Exacerbated by the fact that our software runs on everything from laptops to small clusters to cloud/HPC providers
  - Domestically and internationally
- Opportunity?: NSF Scientific Middleware
  - High performance, including I/O
  - Platform diversity & evolution
  - Assurance of maintenance over decades
  - Complex science data structures, not just linear algebra

www.nrao.edu
science.nrao.edu

The National Radio Astronomy Observatory is a facility of the National Science Foundation operated under cooperative agreement by Associated Universities, Inc.