NRAO

National Radio Astronomy Observatory

Atacama Large Millimeter/submillimeter Array
Expanded Very Large Array
Robert C. Byrd Green Bank Telescope
Very Long Baseline Array

NAASC data processing capabilities and reprocessing scope

Mark Lacy
North American ALMA Science Center

Atacama Large Millimeter/submillimeter Array
Expanded Very Large Array
Robert C. Byrd Green Bank Telescope
Very Long Baseline Array

ANASAC meeting Oct 12-14 2011

NRAO

Outline

- Cycle 0 data delivery
- NAASC support for reprocessing during Cycle 0
- Software status
  - CASA and Pipeline
  - Splatalogue
  - Simdata
  - Future tools and software development meeting
- Computing hardware testing
- Advanced analysis tools

Data transfer to North America

- We will attempt to take advantage of improved links to Chile required by NOAO for DES and LSST.
  - Have agreement for 10% of 1Gb/s link from Oct 1 (now [>]10% of 100Mb/s [1TB/day])
  - Upgrade to 1Gb/s in future (formal AUI/AURA agreement)
- Thereafter data travels via I2/NLR to Charlottesville/UVa
- Should be adequate to move both bulk data and metadata without requiring shipping of disks.
- Archive replication from SCO to NAASC now working successfully. Database synced every 10min.

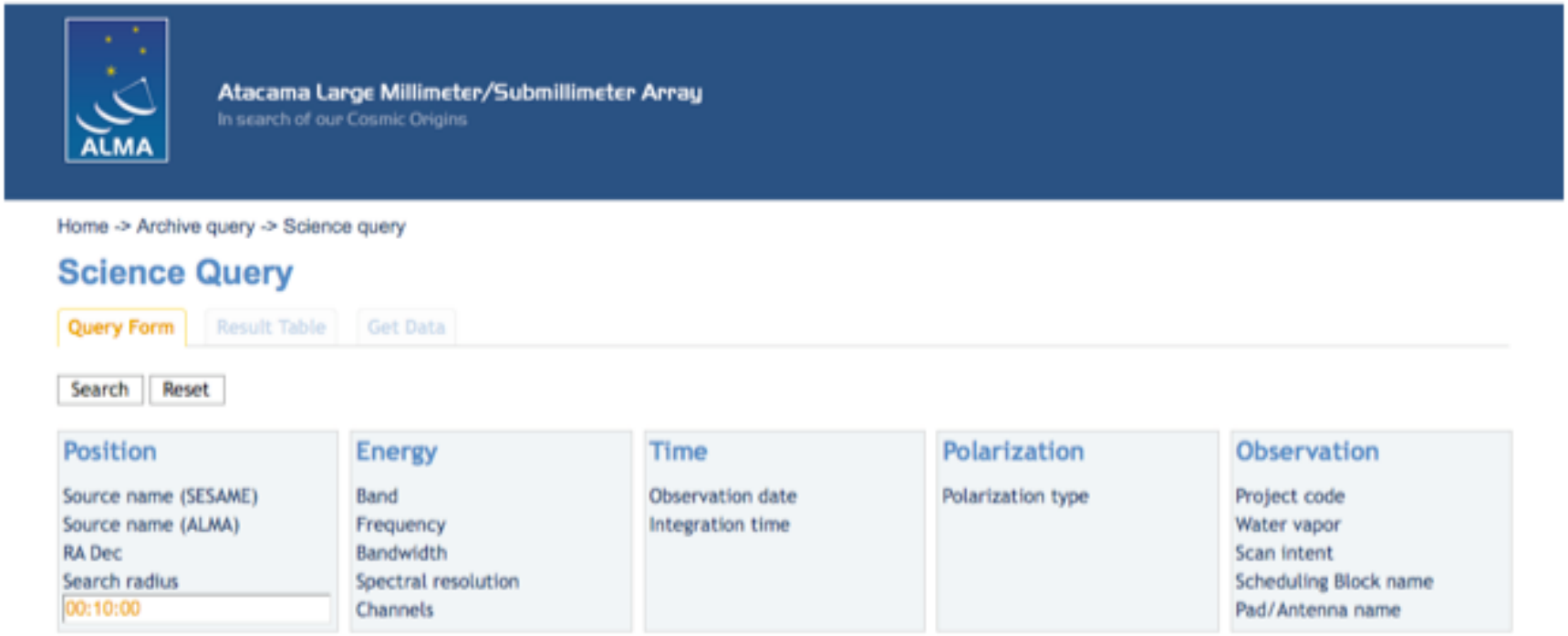
Delivery of Cycle 0 data

- The ALMA Science Archive is still under development. Unlikely to have a "full service" interface and data delivery system ("Request Handler and Data Packer) until June 2012.
- The interim plan is as follows:
  - Data products from by-hand reductions at SCO will be written to a tarball and placed in the archive as a single file in NGAS.
  - This tarball is then replicated to the ARCs.
  - Data analysts will extract the tarballs from NGAS, notify the contact scientist (CS) of data availability. CS verifies data integrity, notifies the PI and starts the proprietary clock.
  - Most data will be delivered via network or USB drive, depending on the PI's institutional network.

ASA status and Archiving of Cycle 0 products

- ALMA science archive will have relational database structure. Major software components (being built at ESO):
  - "metadata harvester" – runs on ASDMs and pipeline products as they are archived, fills fields in relational database.
  - Interface and middleware (VO compliant, based on CADC software)
  - Request handler and data packer (based on ESO software)
  - Expected delivery to users ~ June 2012, individual components may be available before this.
- Current plan is to run pipeline of Cycle 0 when pipeline completed and store these products in the ASA. Tarballs produced by hand will be deprecated.
- Incorporation of user processed data (and QA on it) still being discussed. Most likely will not be stored in ASA proper, but institutional archives may store make available in ASA interface via VO protocols.

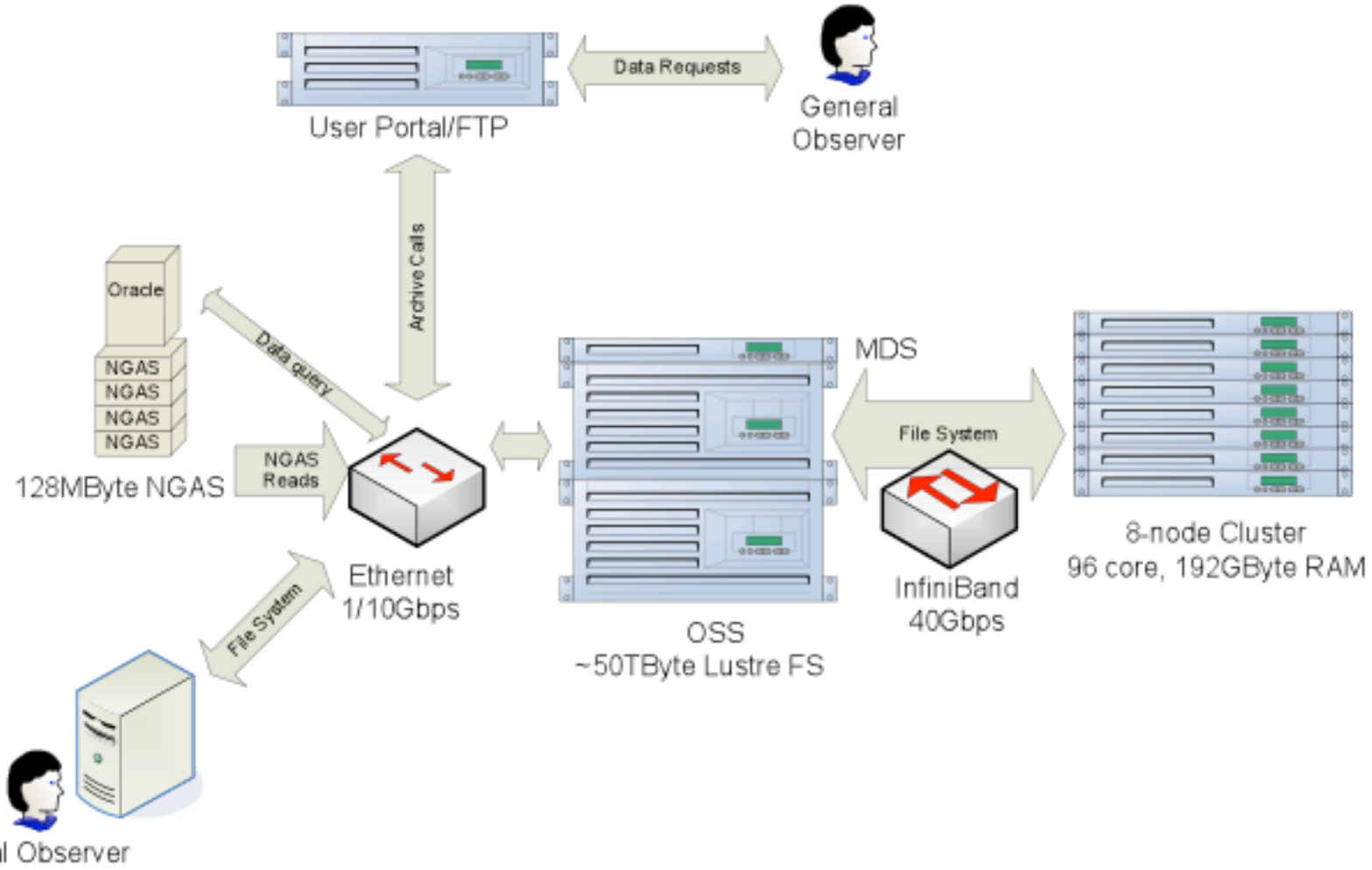
ASA query interface (mock up)



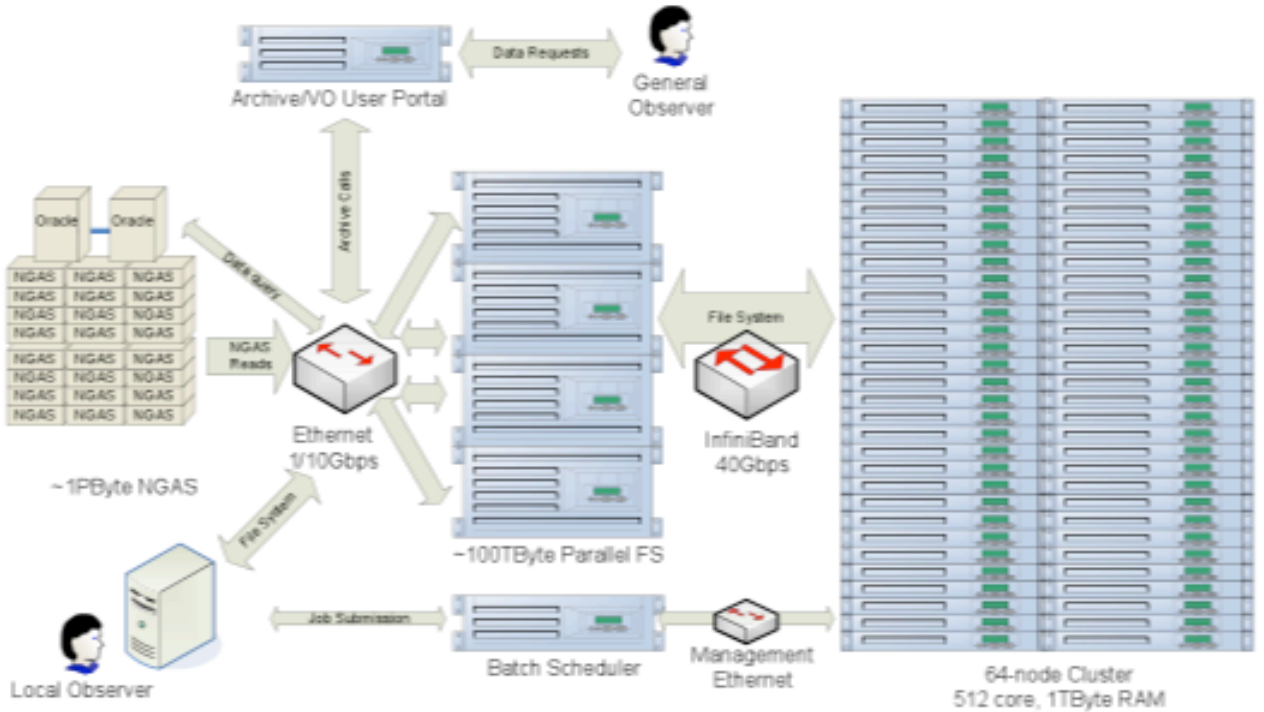
Details of NAASC reprocessing capabilities in Cycle 0

- We have an 8-node cluster in place, connected to a fast Lustre filesytem. In the process of increasing this to a 24-node cluster.
  - We will make this cluster available to visitors.
  - In addition, we have purchased desktop machines for visitor use and for performance evaluation on ALMA data.
  - We will have the compute resources to perform reprocessing of all NA Cycle 0 data at the NAASC if required.

NAASC cluster – current



Full Science Cluster



- Note that growth of cluster will in practice be driven by demand. Slow buildup also allows maximum flexibility (e.g. GPUs or more memory for some nodes) and value for money.

User reprocessing at the NAASC in Cycle 0

NAASC

- The pipeline due date isn't until Q3 2012
- Project will deliver calibrated uv-data (and associated calibration tables) and QA images for cycle 0.
- Users will need to produce publication-quality images (with our help as needed).
  - NAASC visit (preferred). Visitors will have an assigned desktop and/or cluster node. Visitors will be supported by their contact scientist.
  - Home processing. PIs will process their data at their home institute. Support will be through helpdesk and contact scientist.
- This model will change once the pipeline is producing science quality images
  - web interface to the pipeline will allow remote pipeline execution at the NAASC .

NRAO

Pipeline

- The ALMA pipeline development merged with EVLA pipeline and CASA management under Jeff Kern.
- Will enable leveraging of CASA development and expertise
- Pipeline now being made modular, and capable of taking parameters from both users and/or from heuristics algorithms.
- Ed Fomalont is leading Pipeline Algorithms Users Group (PAUG), Remy Indebetouw also a member. This group will provide science oversight to the ALMA pipeline.
- Testing/proofing of the pipeline on Cycle0 data will proceed over the coming year

CASA

NAASC

- CASA 3.3 release @Oct 15 (for Cycle 0); 3.4 @April 15
  - Additionally, monthly "stable" builds for quicker access to new functionality (at the cost of less thorough regression testing and poorer documentation of new features)
  - ~1000 CASA downloads in past 6 months
- NRAO User forum will open 1st Oct (mostly for CASA)
  - Idea is that it will be "self-help" for users.
- CASA questions currently being fielded through NRAO helpdesk, expect some through the ALMA helpdesk too.
- CASA prioritization for ALMA related development lead by CASA subsystem scientist for interferometry (C. Brogan) and Single Dish (D. Iono) with input from ARCs, JAO, and feedback from users at workshops and helpdesk

NRAO

Three main areas of CASA focus in coming year:

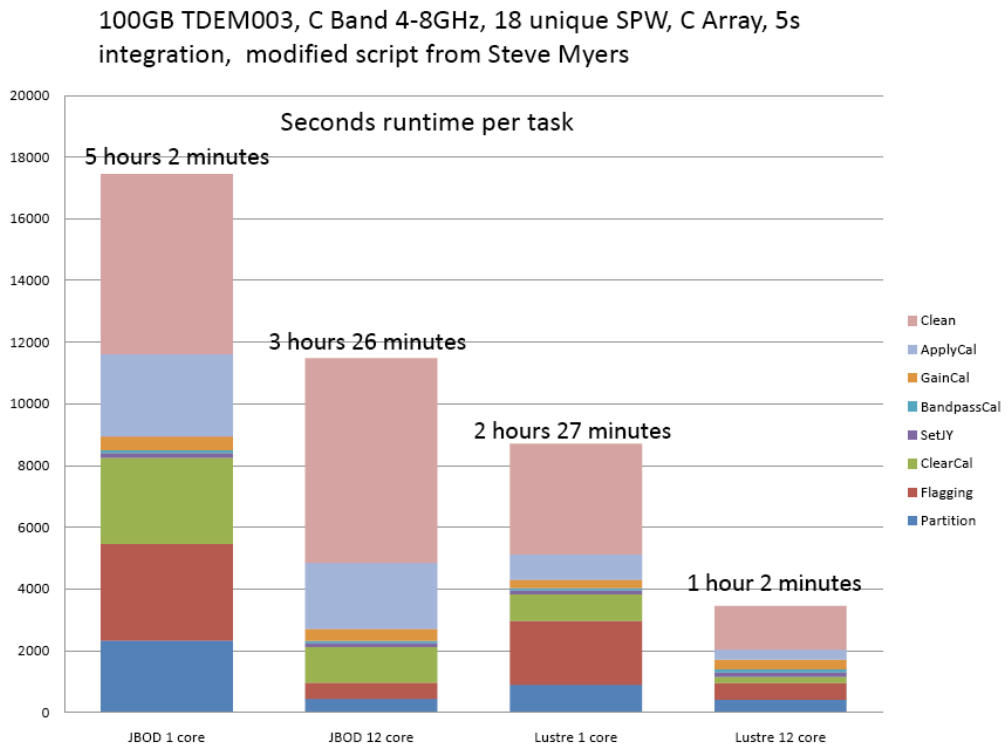
- Parallelization
- Completing suite of calibration and imaging needs through full science
  - Polarization
  - Joint deconvolution for single dish and interferometric data
  - Band to band calibration transfer (Band 6 to Band 9 for example)
- Adding more *basic* visualization and analysis functionality
  - These upgrades will facilitate "zeroth" order visualization and analysis AND ability to hook into more advanced data analysis tools
  - Search for new hire in this area underway
- Also through NRAO Algorithm Research and Development Group (tackles research level problems across observatory):
  - Autoflag, antenna dependent pointing calibration, and spectral index dependent continuum imaging

CASA parallelization

- CASA uses an "embarrassingly parallel" approach.
  - *Partition* task has been added to form smaller MSs, a reference MS implementation simplifies the user view
- Parallelized:
  - *clearcal* and *applycal* tasks are parallel
  - *imaging* has been parallelized at the tool level.
- To Do (release 3.4+):
  - Additional Tasks: hanningsmooth, setjy, flagging, filler
  - Task based imaging
  - Robustness and usability
  - User Education
- Pipeline parallelization can take place at three levels:
  - Within a CASA command using the parallel CASA infrastructure.
  - By running multiple "Job Data" objects simultaneously
  - By running multiple pipelines simultaneously
    - Scheduling done external to pipeline and CASA

Computing hardware testing

- HPC (Robnett) – detailed breakdown of i/o, CPU, memory to optimize cluster.
- Results show parallel CASA speedup, also speedup with Lustre



Computing hardware testing (https://science.nrao.edu/facilities/alma/naasc-hardware-recommendations)

- NAASC level – scripts on single machines ("runtime estimates")
- Preliminary results suggest clock speed seems to win over number of cores and i/o not rate limiting for non-interactive tasks (though interactive tasks are faster on the Lustre filesystem) – situation expected to change drastically when CASA parallel is available.

System	Antennae Band 7 Calibration	Antennae Band 7 Imaging	Antennae Band 7 both
System 1	3.7 hr	0.6 hr	4.3 hr
System 2	2.3 hr	0.5 hr	2.8 hr
System 3	3.0 hr	0.7 hr	3.7 hr
Single cluster node	3.0 hr	0.7 hr	3.7 hr

•System 1: Dual core 2.27GHz , 12GB RAM, local (non-RAID) disk
•System 2: Quad core 2.8GHz, 24GB RAM, RAID0 disk
•System 3: Dual quad core 2.26GHz, 24GB RAM, RAIDed disk
•Cluster node: dual hex core 2.26GHz, 24GB RAM. Lustre disk

NAASC and related software systems

- Splatalogue
  - Currently concentrating on documentation and database enhancement.
  - Future plans include improvements to usability (new front end). "Quick picker" in place already.
  - International working group set up (currently addressing lines at >1THz)
  - Subset of Splatalogue now in CASA, also new CASA task to read output from Splatalogue into a CASA table
- Simdata (task in CASA)
  - Simdata now largely complete, including single dish capability (in collaboration with NAOJ).
  - Split into simobs and simanalysis tasks in 3.3
  - Simdata helped us to select Cycle 0 configurations, and enabled us to write memo on the effects of clean bias on ES images.

NAASC advanced tools

- A key role of the NAASC is to ensure that users can successfully analyse the large data cubes ALMA will produce
- We hope to get support from the community to develop advanced analysis tools via the ALMA development program
  - ALMA Software Development workshop Oct 12-14 in Charlottesville to bring together interested community members
  - Our new hire in this area will also allow us to initiate some efforts ourselves as well as linking community efforts to CASA