Increasing the ALMA data rate

Mark Lacy

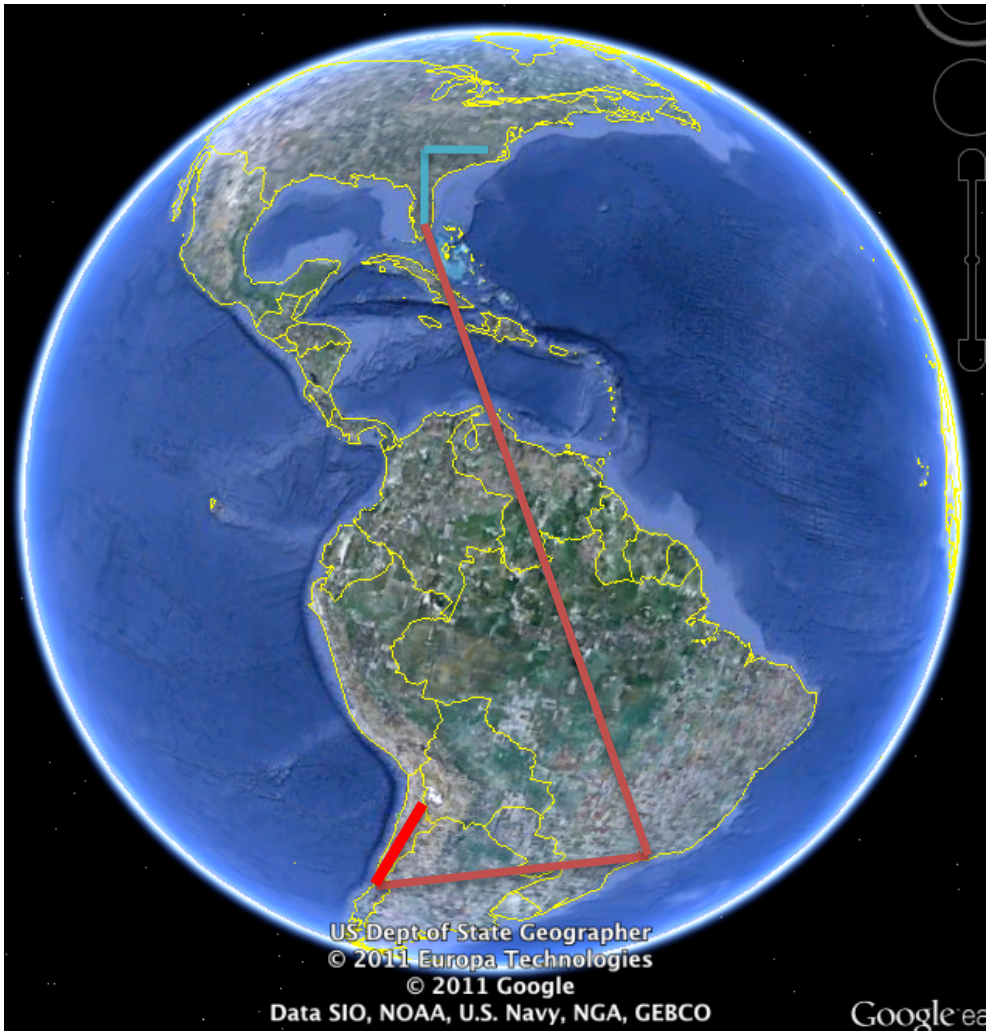
Data Services Lead, NAASC, NRAO

Some science cases for high(er) data rates and their implications

- Larger bandwidths
  - More continuum sensitivity.
  - More lines (at the same resolution) per observation.
  - *Probably ~OK with scaling of existing data transfer infrastructure for up to ~32GHz bandwidth.*
- New observing techniques
  - On the fly interferometry (fast surveys)
  - *Probably ~OK with scaling of existing raw data transfer infrastructure, major issue will be imaging.*
- Focal plane arrays
  - Potential huge increase in survey speed
  - *Depending on size of arrays (and if they are installed on all antennas, not just TP and/or ACA), may need a dramatic change in the data management plan.*

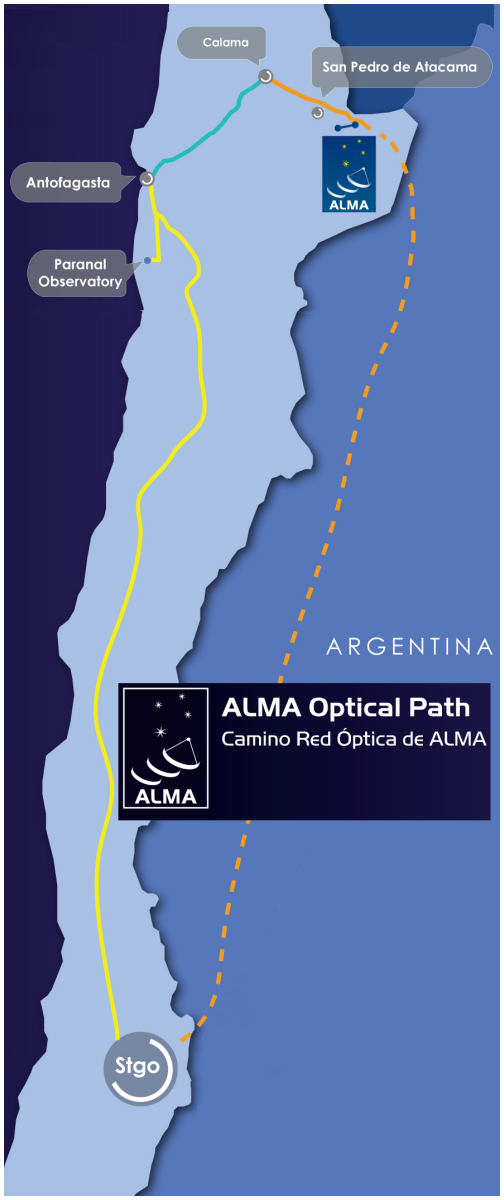
Data transfer

- Three stages:
  - AOS to Santiago
  - Santiago to Miami
  - Miami to Cville



Data transfer within Chile

- Upgrade project from ALMA development program gives 2.5Gb/s from AOS to Santiago:
  - OSF to Calama fiber built 2014; waiting on revised environmental impact report before “official” use, unofficially is now working.
  - Calama to Antofagasta provided by Telefonica
  - Antofagasta to SCO from EVALSO/REUNA (Chilean academic network provider)
  - Redundant fiber loop via Argentina planned
- Primary ALMA archive in Santiago (SCO)
- Santiago to ARCs: individual ARC contracts.



Santiago to Miami

- NRAO works with the South American Astronomy Coordinating Committee (SAACC) to provide network to the US.
- Joint AUI-AURA agreement gives NRAO 100Mb/s burstable to 600Mb/s (in practice). Can be improved if/when needed.
  - Main provider is Amlight (Florida International University).
- Also AUI-REUNA MOU for local transfer from ESO campus to hub in Santiago.
- Network links to South America improving rapidly.

Within the US

- Within the US, use academic high speed networks (Internet 2) to University of Virginia.
- Once at UVa, 2Gb/s link to NRAO (will upgrade to 10Gb/s).

Current main bottleneck is thus Santiago to Miami, can be improved if needed/justified though.

Current ALMA data rates

- Operations plan D assumed 200TB/yr (6.3MB/s) in Full Science.
  - Early (Cycle-1/2) fears that this data rate would be greatly exceeded have not been founded (helped by Phase-2 policies and user education), and data rate justification has been removed from proposals.
- ALMA in Cycle 4 will produce about 100TB (/yr) in raw data.
  - (in practice 200TB/yr as data will be stored both WVR corrected and uncorrected, but this should only be temporary)
  - Image products are currently only ~10% of data, but this is expected to increase significantly when the imaging pipeline is fully operational.
- For **raw** data in Full Science (more efficient plus a few more antennas than Cycle 4), the ops plan estimate is probably good, will need to increased depending on the size of the products (largely dictated by processing resources). We are assuming 500TB/yr in Full Science operations (Cycle 5+).

“Hard” data rate limits

- Correlator network infrastructure (64MB/s [512Mb/s])
  - Low enough that for some projects (long baseline [short sampling time]) and full resolution spw we hit this limit (especially when taking both WVR corrected and uncorrected data streams).
  - No problem getting this to Santiago over 2.5Gb/s link.
  - Could (fairly) easily boost SCO->MIA link to this capacity.
  - An improvement would allow better long baseline observations, and a richer archive.
- Raw correlator output is 512MB/s (using 4-bit visibilities). Would be difficult to transmit (4Gb/s).

The ALMA data challenge

- Existing infrastructure can probably support raw data rates ~2-4 times larger than at present. (Other ARCs not quite so well situated, but solutions could be found.)
- Processing the data can be a challenge though. Currently 2.5 months behind on processing (dominated by organizational issues).
  - Until now, reference images only generated. Pipeline into operation in Cycle 4.
  - Imaging demands depend on configuration (problem scales with longest baseline squared).
    - Small configurations (<1km), image data volume <~ raw data volume even if mapping all channels at full resolution to the edge of the primary beam.
    - Large configurations more problematic, image data volume can greatly exceed raw for short snapshots, imaging process can run for weeks (radio interferometry can be a very efficient compression algorithm!).
    - Still have to explore what we can practically make for pipeline image products in large configurations.
    - Images also need to be mirrored out of SCO master archive, increasing load on data transfer out of Chile

Summary

- Strong science cases exist for increased data rates arising from improvements to correlators and receivers.
- An agreement with AURA (DES, LSST) have allowed NA to obtain good data connections to Chile. Improvements in connectivity to South America (triggered by Rio Olympics) mean that regular internet is also much better (and can be used for backup).
- A factor of 2-4 increase in raw data rate could probably be supported at a reasonable cost (though would need to be justified and accounted).
- Still uncertain is the cost/difficulty of imaging even the current data in the largest configurations at full spatial+spectral resolution over the full primary beam, and its implications for data transfer to/from Chile.