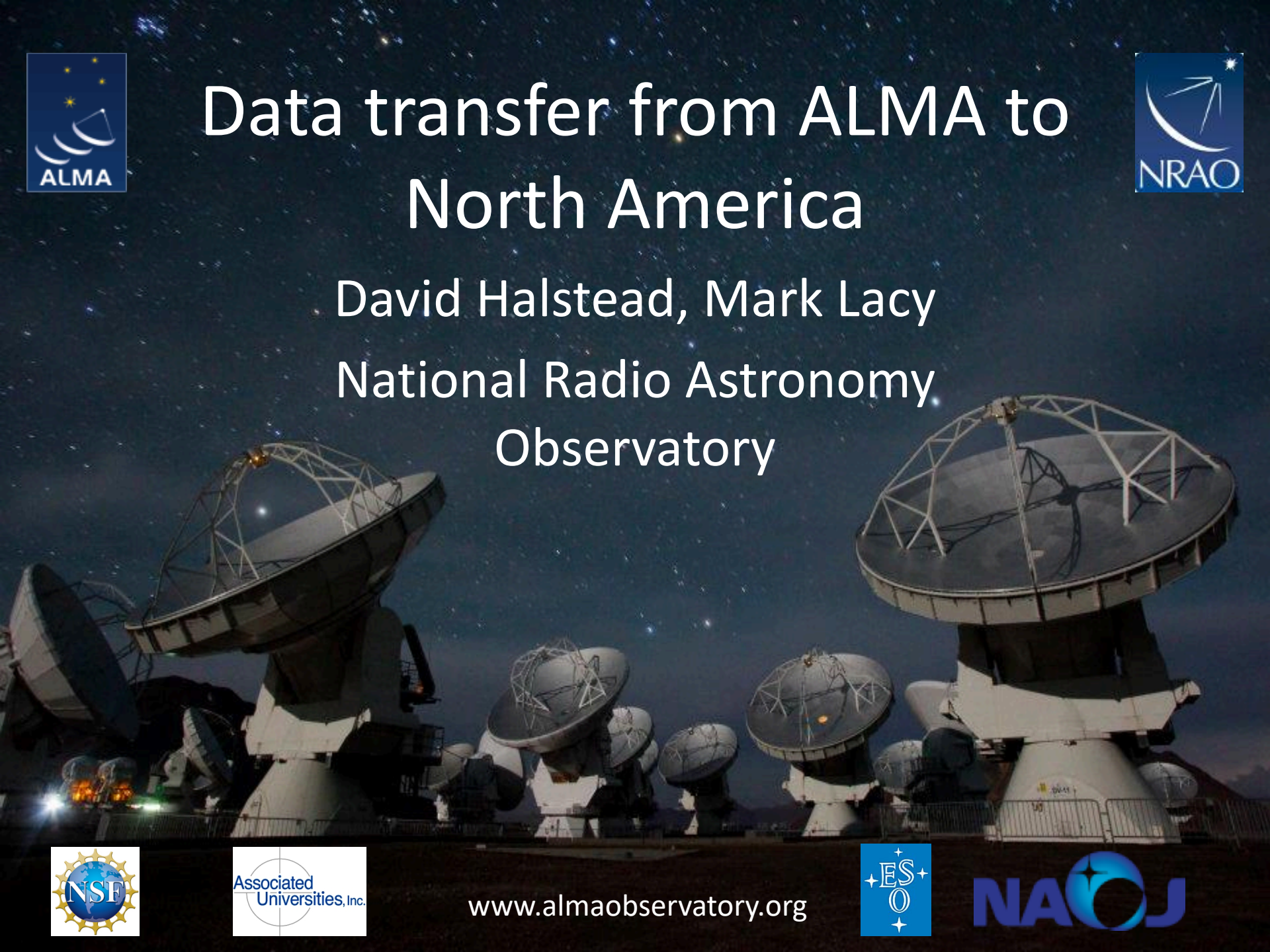
# Data transfer from ALMA to North America

David Halstead, Mark Lacy

National Radio Astronomy Observatory



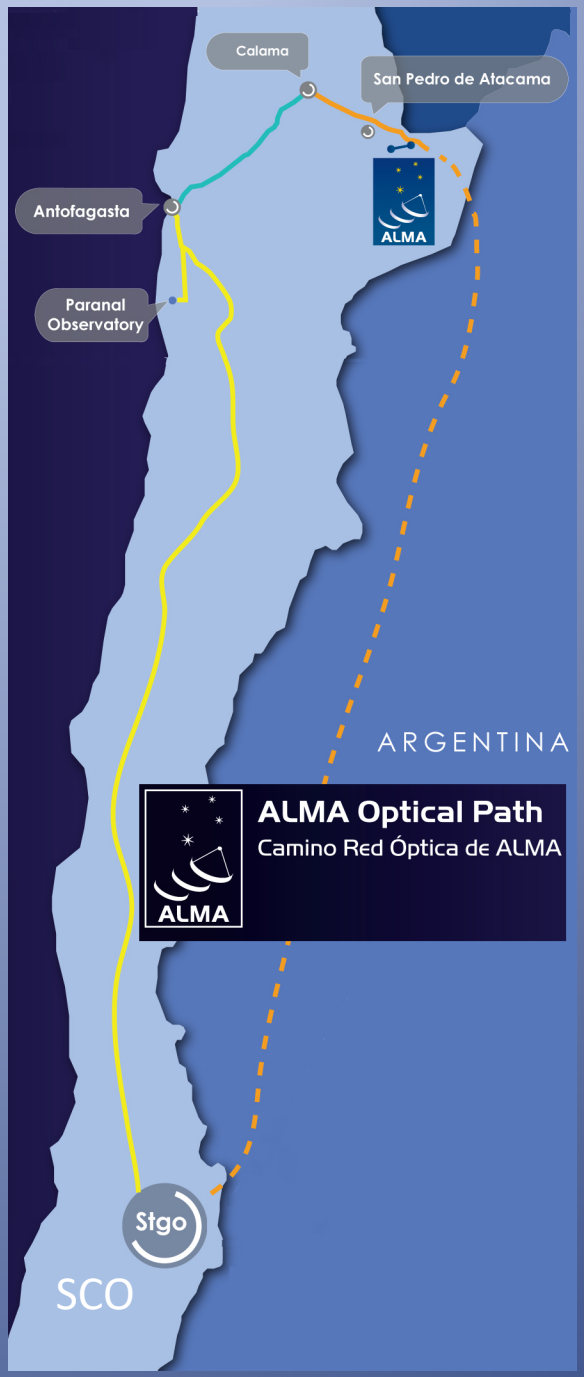
www.almaobservatory.org

# ALMA telescope

- Largest mm/submm telescope ever built.
- Interferometer – combines signals from multiple antennas to form an image.
- All 66 antennas delivered, all at high site (except for maintenance).
- Multinational project with many partners, three ALMA Regional Centers (ARCs): US, EU and EA
- Operated “space mission” style, with pipeline data processing and a science archive at each ARC allowing data reuse.
- First PI projects released to public from the ARCs January 2013
- Cycle 5 observations began in October 2017.

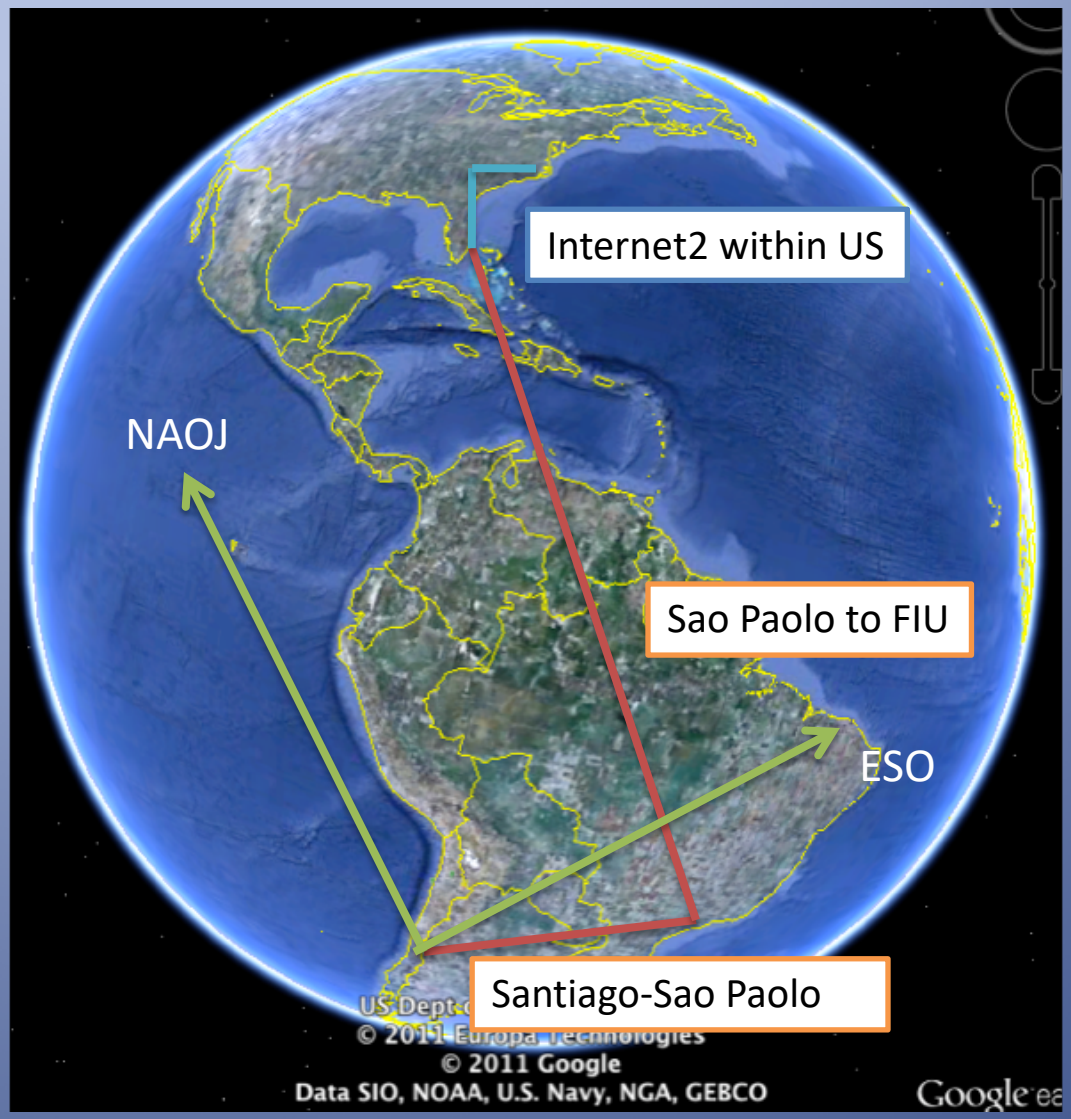
# Data Transfer within Chile

- AOS to Santiago 2.5Gb/s (fiber to Calama, commercial fiber Calama to Antofagasta, EVALSO/REUNA from Antofagasta to Santiago.
  - Redundant fiber loop via Argentina planned.
- Santiago to ARCs: individual ARC contracts with REUNA and NRENs
- Data processing to produce Level 2/3 products shared between Santiago and the ALMA regional centers.
  - Pipeline is run at 4 locations worldwide, including Santiago.
  - Data packages ingested into the archive in Santiago.
  - Pipeline products ~ same size as raw data.
  - A significant amount of data (~100TB/yr) is thus being uploaded from the NA ARC to JAO. Bi-directional speed is thus important.
- Long-term plan is that all data processing will take place in Santiago.

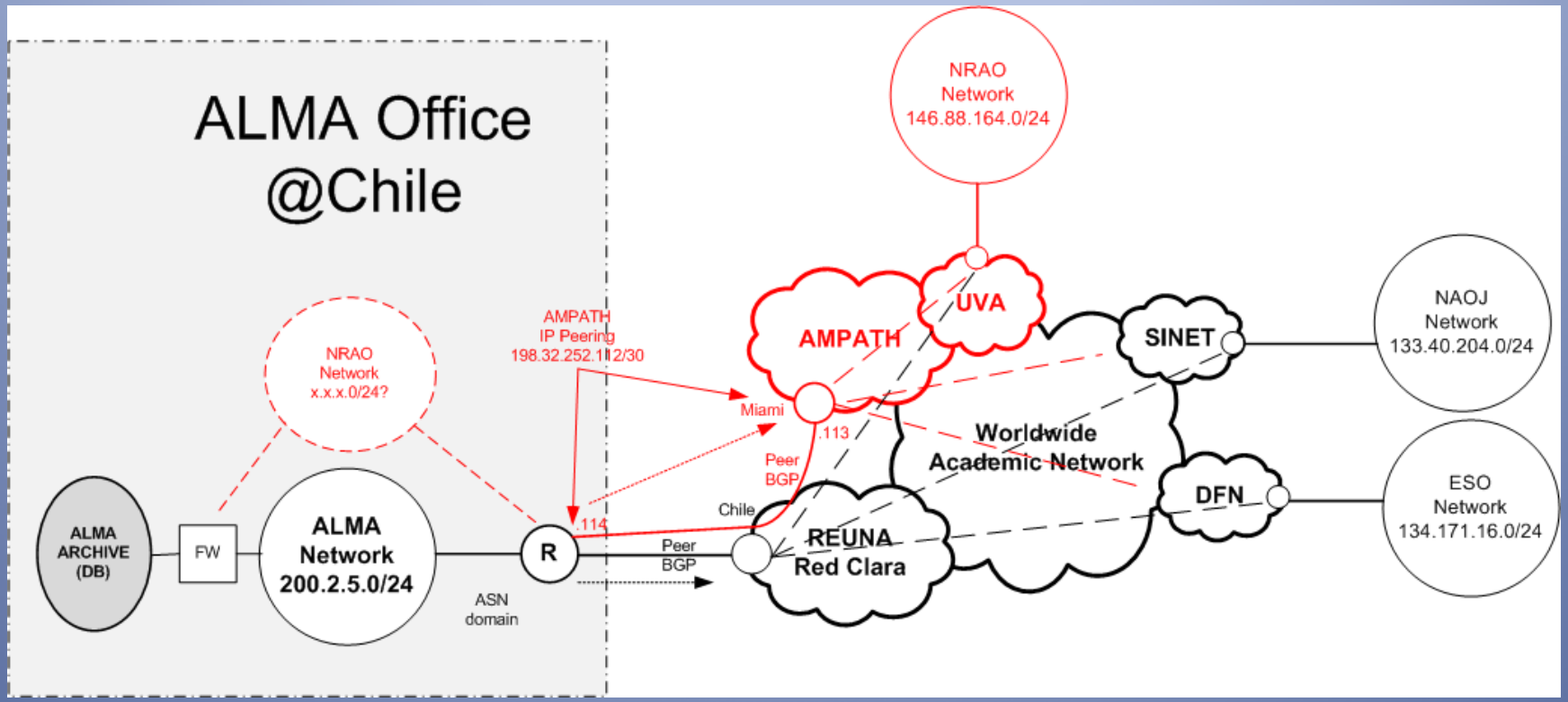


# Data transfer – Chile to NA

- Joint AURA-AUI agreement for 100Mb/s committed (burstable to capacity) of AURA's link to Chile through Sao Paolo and Miami (FIU/AmLight) to the US research network backbone (NREN).
- MOU signed between AUI/REUNA for local link to SCO.
- Link from NRAO to Internet2 through UVa is 10Gb/s
- Typical rate obtained during peak data transfer periods is 2-300Mb/s, with bursts up to 600Mb/s.
- Currently working on establishing network monitoring, and improving our understanding of how the link performs in typical load conditions (~1TB/day).



# Paths from ALMA



# ALMA Science data rate evolution

- ALMA Cycle 0 completed (Oct 2011-Jan2013)
  - 16-24/50 antennas used (data rate proportional to square of antenna number)
  - ~5-10% of array time for science
  - Total data volume was about 20TB
- ALMA Cycle 1 complete (~Aug 2013-Jun2014)
  - 32-40/50 antennas, plus 7/12 compact array
  - ~10% of array time for science
  - 40TB over 1yr (ALMA archive hit 50TB in March 2014)
- ALMA Cycle 2 complete (June 2014-Sept 2015)
  - ~34 main array antennas, 10 compact array
  - ~15% of array time for science (but some carryover from Cycle 1)
  - 70TB in a 17 month Cycle.
- ALMA Cycle 3 complete (Oct 2015-Sept 2016).
  - 36 main array antennas, 10 compact array
  - ~25% of array time for science
  - Total of 140TB, mostly raw data (manual imaging and imaging pipeline products only ~20%).
  - Data volume artificially high as two data streams are kept with different corrections.
- ALMA Cycle 4 complete (Oct 2016-Sept 2017)
  - 40 main array antennas, 10 compact array
  - ~33% of array time for science
  - Total of 210 TB, mix of raw and pipeline image products
  - Data volume artificially high as two data streams are kept with different corrections.

# Cycle 5

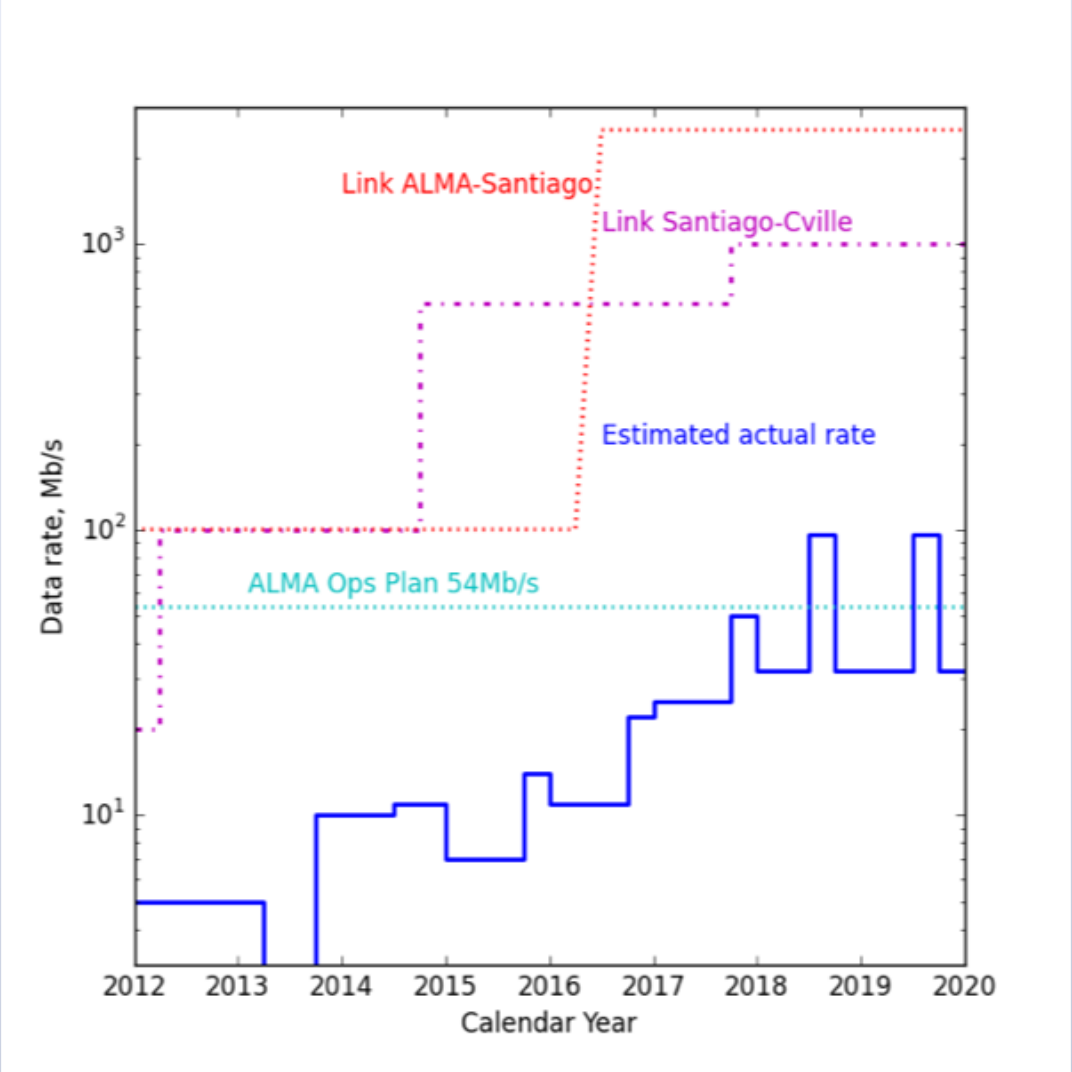
- Still taking data in 2 streams (WVR corrected and uncorrected)
- ~45% of array time for science
- More antennas (43 compared to 40).
- Total volume will be ~360TB, including image products (which will constitute about 30% of the total data volume).

# Future Cycles

- Expect Full Science cycles (Cycle 6 onwards) to have mean data rates ~100Mb/s during observations. “Duty cycle” of observations will slowly increase as testing and maintenance procedures improve.
- WVR-corrected raw data stream will probably be turned off in October, reducing raw data volume by a factor of two, and total by ~30%.
- Best guess estimate for the next 5 years (including product size mitigation) is around 250TB/yr (125TB raw, 125TB products).
- Important to note that data rates vary through the configuration cycle. When long baseline configurations are scheduled the data rate goes up for two reasons:
  - Data sampling needs to be faster to prevent beam smearing at the field edges.
  - The data products, which are also mirrored from Santiago, also increase in size, to become larger than the raw data in the largest configurations.
  - So far, long baseline campaigns have tended to have low observing efficiencies, however this may change.

# Current data rate projections

- Assumes no imposed limit on data rate (cyan line is current Operations Plan rate).
- Blue line is for data generation
- Data transmission is per ARC



# Correlator upgrade

- A correlator upgrade is scheduled for 2022.
- This will allow up to 8x more channels, and a 2x wider bandwidth.
- Expected data rate increase is about a factor of four, corresponding to a data rate of 1PB/yr (not all projects will need the extra channels, though most will use the wider bandwidth).
- Some increase in data rate can be expected in the build up to this as the the network at the AOS is upgraded (but should only be modest, ~10% overall).

# Summary

- Ramp-up of the ALMA data rate has been slower than anticipated, allowing us to stay ahead of the curve.
- 2-stream data collection in Cycles 3-5 is artificially boosting the data rate, we are assuming this will no longer be the case in Cycle 6 (Oct 2018 onwards).
- Bidirectional data flow will continue to be needed to support data processing at the ARCs.
- Still learning how the network performs when transferring ~1TB/day in multiple parallel streams.
- Would like to establish a link with 1Gb/s available bandwidth (out of a 10Gb/s pipe) within the next 1-2 years to improve our transfer speed to and from Chile for bulk reprocessing, and to help with occasional large data and metadata transports (e.g. a DB export).
- New developments on ~5yr timescale (e.g. the correlator upgrade) can probably be accommodated with a modest (x4) increase the data rate.