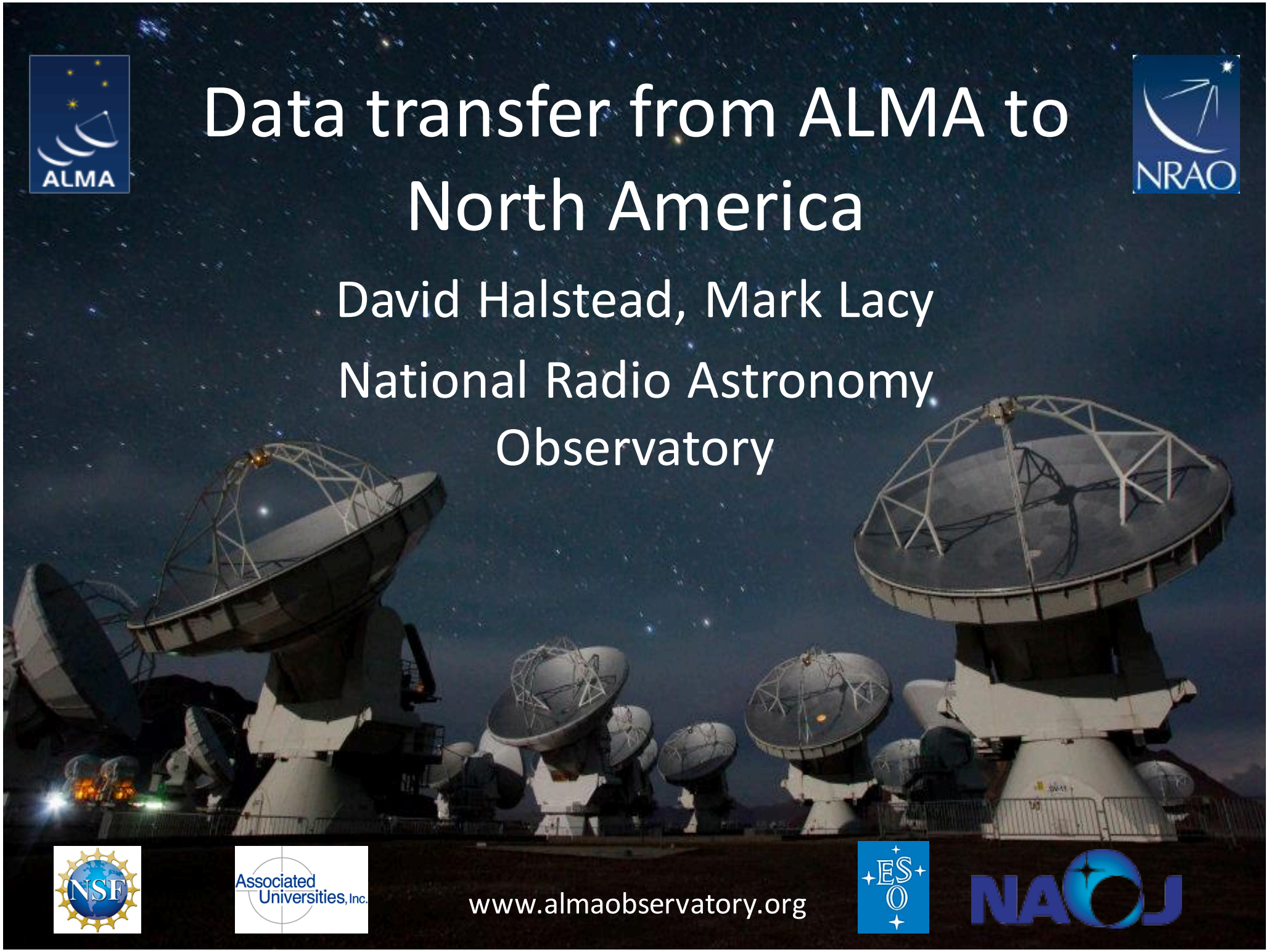
Data transfer from ALMA to North America

David Halstead, Mark Lacy

National Radio Astronomy Observatory



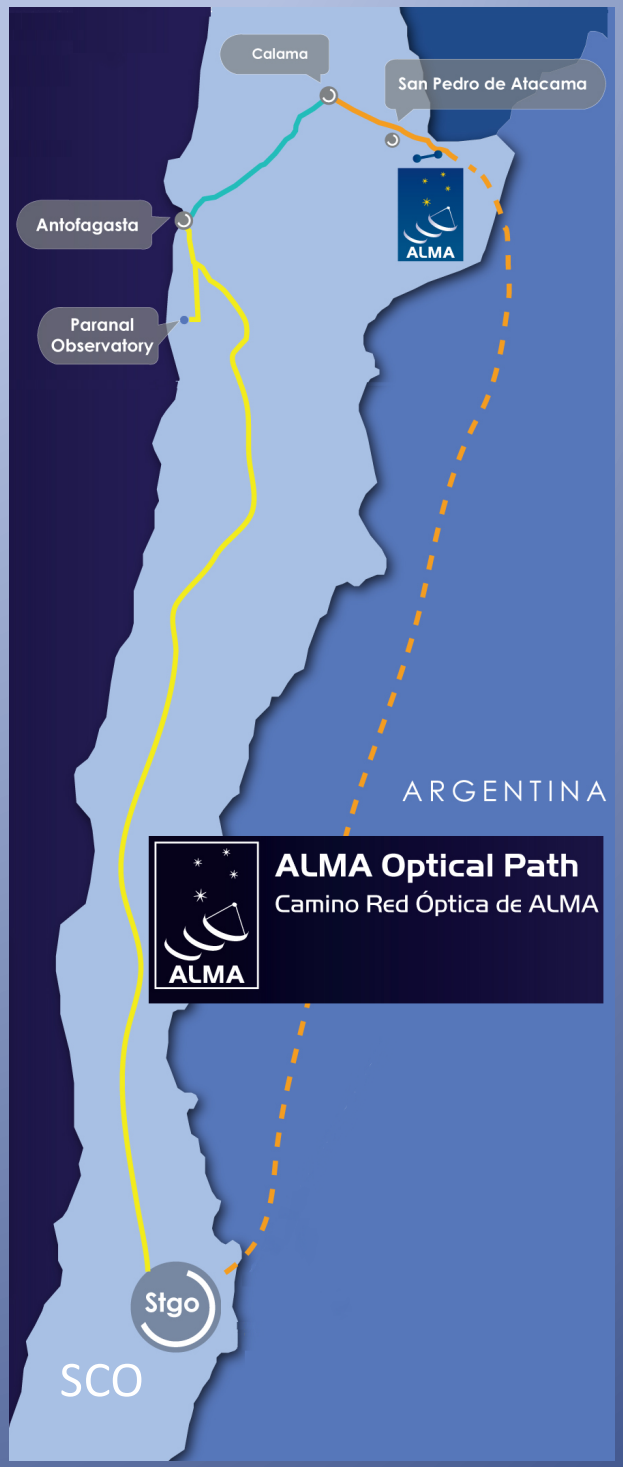
www.almaobservatory.org

ALMA telescope

- Largest mm/submm telescope ever built.
- Interferometer – combines signals from multiple antennas to form an image.
- Inauguration occurred at the OSF on March 13th 2013
- 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 4 observations began in October 2016.

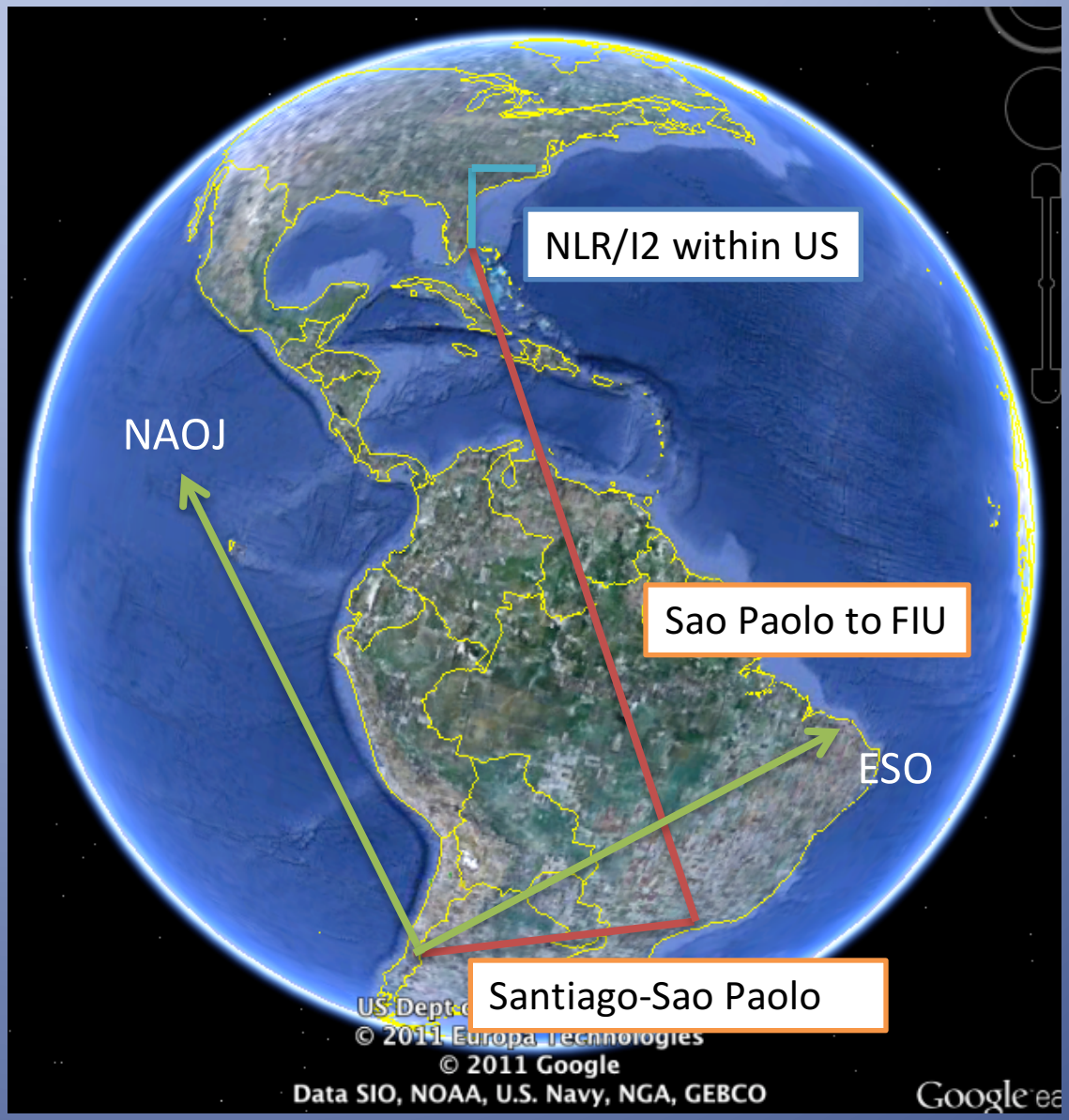
Data Transfer within Chile

- 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.
- 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.
- Long-term plan is that all data processing will take place in Santiago.
- Santiago to ARCs: individual ARC contracts with REUNA and NRENs

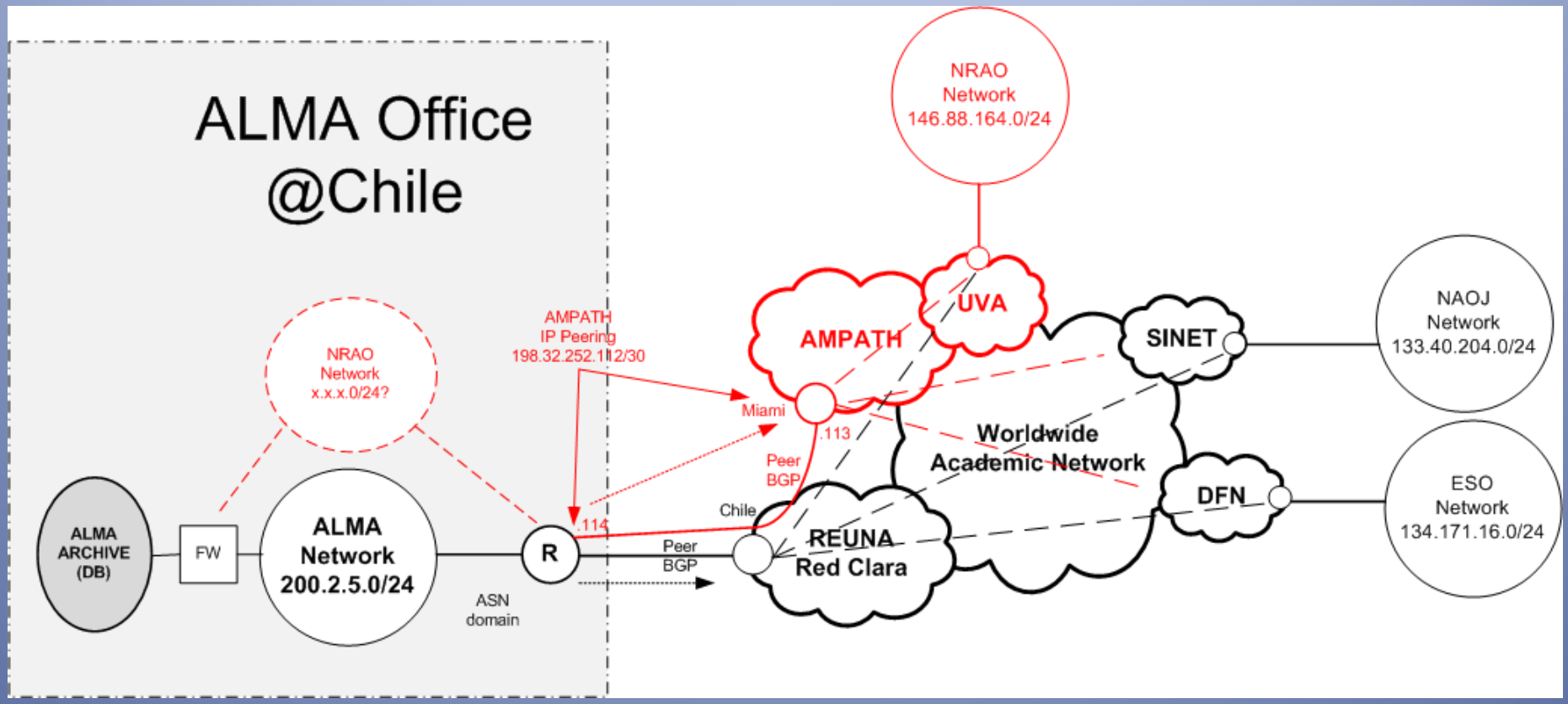


Data transfer – Chile to NA

- Joint AURA-AUI agreement for NRAO to have 100Mb/s committed (burstable to capacity) of AURA's 622Mb/s link to Chile through Sao Paolo and Miami (FIU/AmLight) to the US research network backbone (NREN).
- MOU signed between AUI/REUNA
  - local link to SCO.
  - implementing international links.
- Typical rate obtained during peak data transfer periods is 2-300Mb/s, with bursts up to 600Mb/s.
  - 90% is bulk data with low QoS.
  - Remainder is database sync and telepresence
- MOUs in place between AURA/AUI and AUI/REUNA
- Opportunities for improving Chilean astronomer access under consideration



Paths from ALMA



Note: NRAO will be abandoning 146.88.x.x IP space in the next month

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
  - Data inflated to supply users with intermediate products
  - 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
  - Users will not get intermediate products, better software means unnecessary data not taken.
  - 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.
  - Mean data rate of 40Mb/s during observations.

Cycle 4

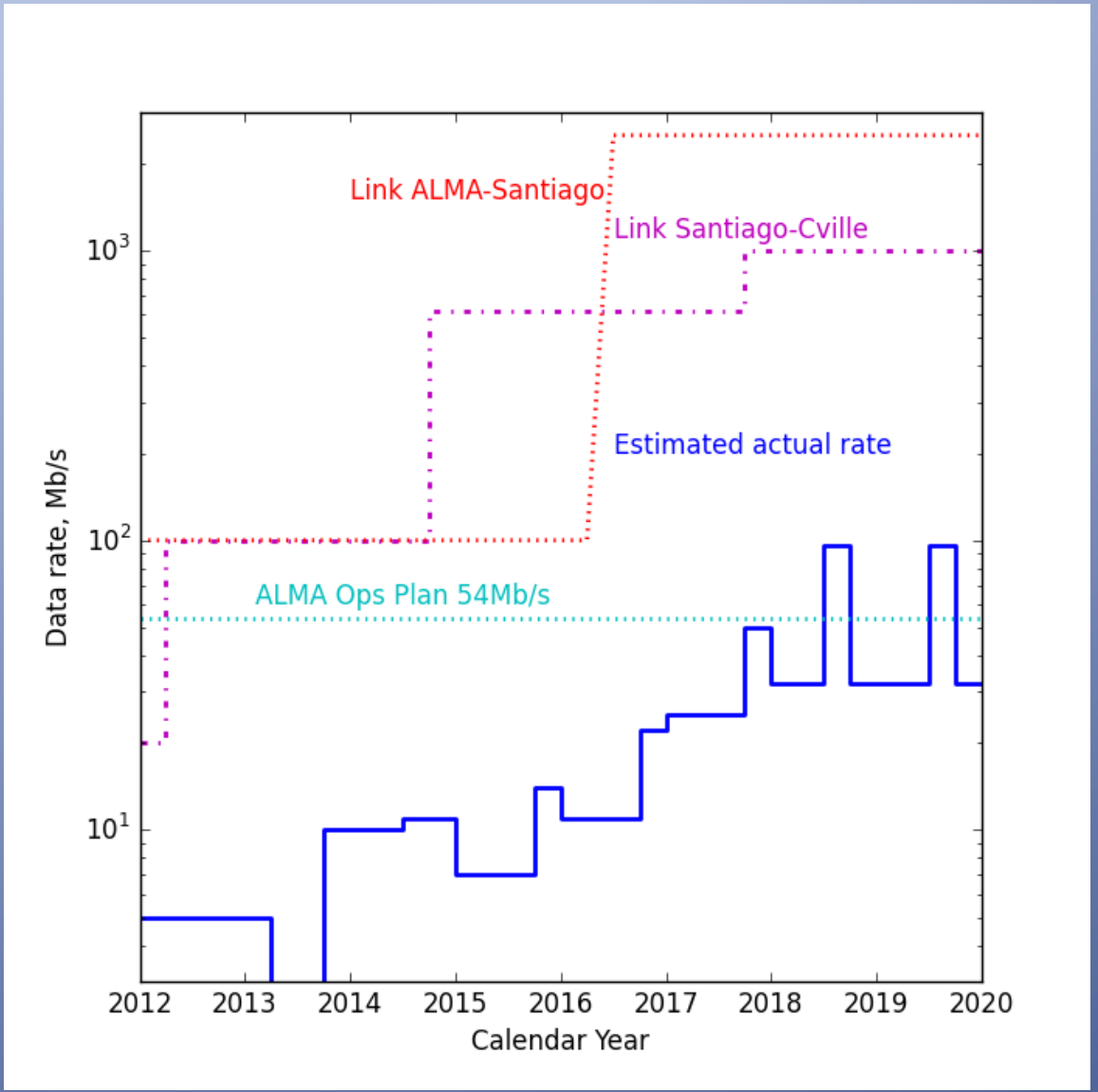
- Still taking data in 2 streams (WVR corrected and uncorrected)
- About the same amount of time available for science on the array, though increased efficiency.
- More antennas.
- Products will be larger than Cycle 3, anticipate will not be full cubes for all sources though.
- Total volume ~190TB, including products.

Future Cycles

- Raw data rates will increase as we transition into full operations. Expect Full Science cycles (~2018 onwards) to have mean data rates ~100Mb/s during observations, but could be ~50% higher. "Duty cycle" of observations will also increase (by about a factor of two) as testing and maintenance procedures improve.
- Product data rates will increase to approximately equal raw data rates once imaging pipeline is creating full cubes for all sources.
- Best guess estimate (including product size mitigation) is around 400TB (200TB raw, 200TB products) (larger if we continue to take 2 streams).
- 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



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+4 is artificially boosting the data rate, we are assuming this will no longer be the case in Cycle 5 (Oct 2017 onwards).
- Full image products will have a large impact in Cycle 5 and beyond, however.
- Most new developments (e.g. correlator improvements) on <10yr timescales can probably be accommodated without increasing the data rate by more than a factor ~2-4.