Archive and Data Services Team

Mark Lacy

ANASAC f2f Meeting – 31 May 2017

NRAO NSF Associated Universities Inc

ALMA Data Reduction

Context

- As the ALMA data rate has increased, the volume of data to be processed through QA2 per cycle has grown.
- It was evident early in Cycle 0 that the original operations plan for data processing that had all data being processed at JAO, was going to need supplemental help from the ARCs.
- ARCs performed manual data reduction for Cycle 0 and Cycle 1 data, based on standard scripts for calibration, and ad-hoc imaging scripts.
- The calibration pipeline was introduced for Cycle 2 (Oct 2014).
- In Cycles 2 and 3 the calibration pipeline was run at JAO, with the ARCs restoring the calibrated products using calibration tables from the JAO.

ALMA Data Reduction

Cycles 2 and 3

- With the calibration pipeline, an extra QA step was introduced, the weblog review.
  - By the end of Cycle 2, a backlog of weblog reviews had built up at JAO. Pipeline runs were being reviewed more slowly than the data was being taken.
  - The NAASC sent DAs to JAO to help clear the backlog, this action alleviated the backlog for Cycle 2.
- The higher data volume in Cycle 3, combined with a slow pipeline deployment, resulted in the weblog review process getting more backed up.
  - The NAASC supplemented the JAO effort by running the pipeline locally for short (1-2 week) periods ("Operation Jaws").
  - Nevertheless, by the summer of 2016, a large backlog of weblog reviews had built up.

ALMA Data Reduction

Cycles 3 and 4

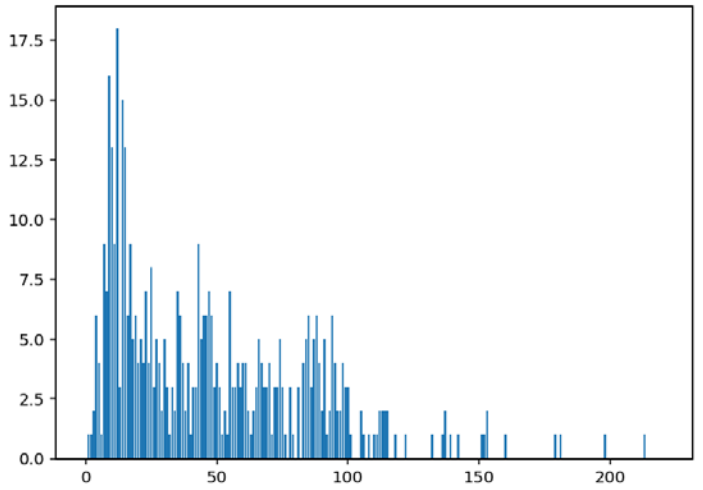
- At the beginning of June 2016, NAASC DAs received permission to remotely review the weblogs at the JAO from Cville.
  - Focussed on NA datasets, but helped out with all.
- For Cycle 4, NA elected to reduce all its own data, both calibration and imaging.
- The full pipeline (including imaging) was made available at the start of Cycle 4.

- Some definitions:
  - "Steady State" – data being processed as fast as it is being taken.
  - "Backlog" - sum of datasets being processed and in queue for processing. Should be only ~20-30 in steady state.

ALMA Data Reduction

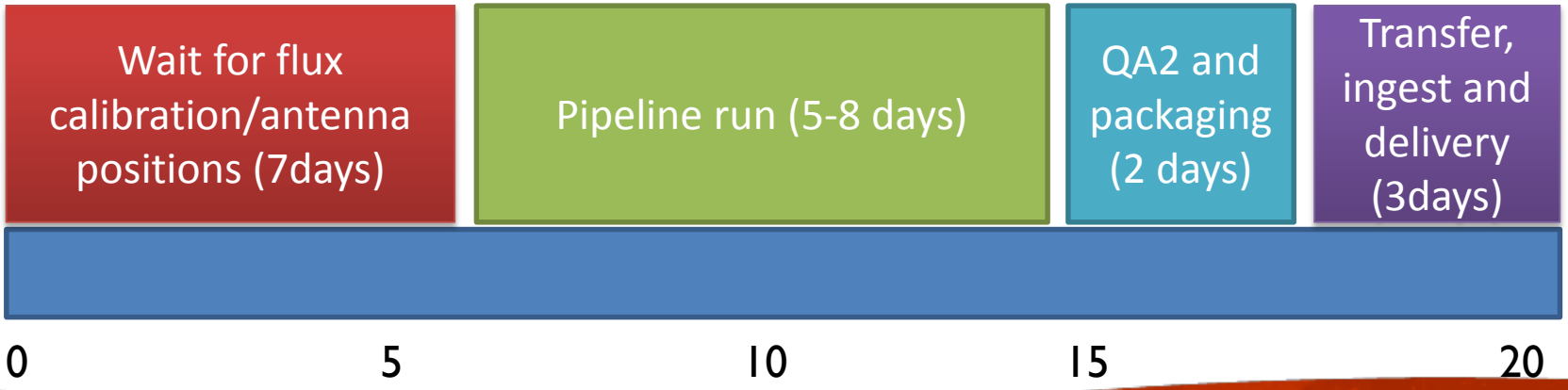
Mean times to delivery

- Most datasets can be delivered ~20d after data obtained (target 30d).
- Manual calibrations, or data requiring significant manual imagin intervention takes longer (target 45d).



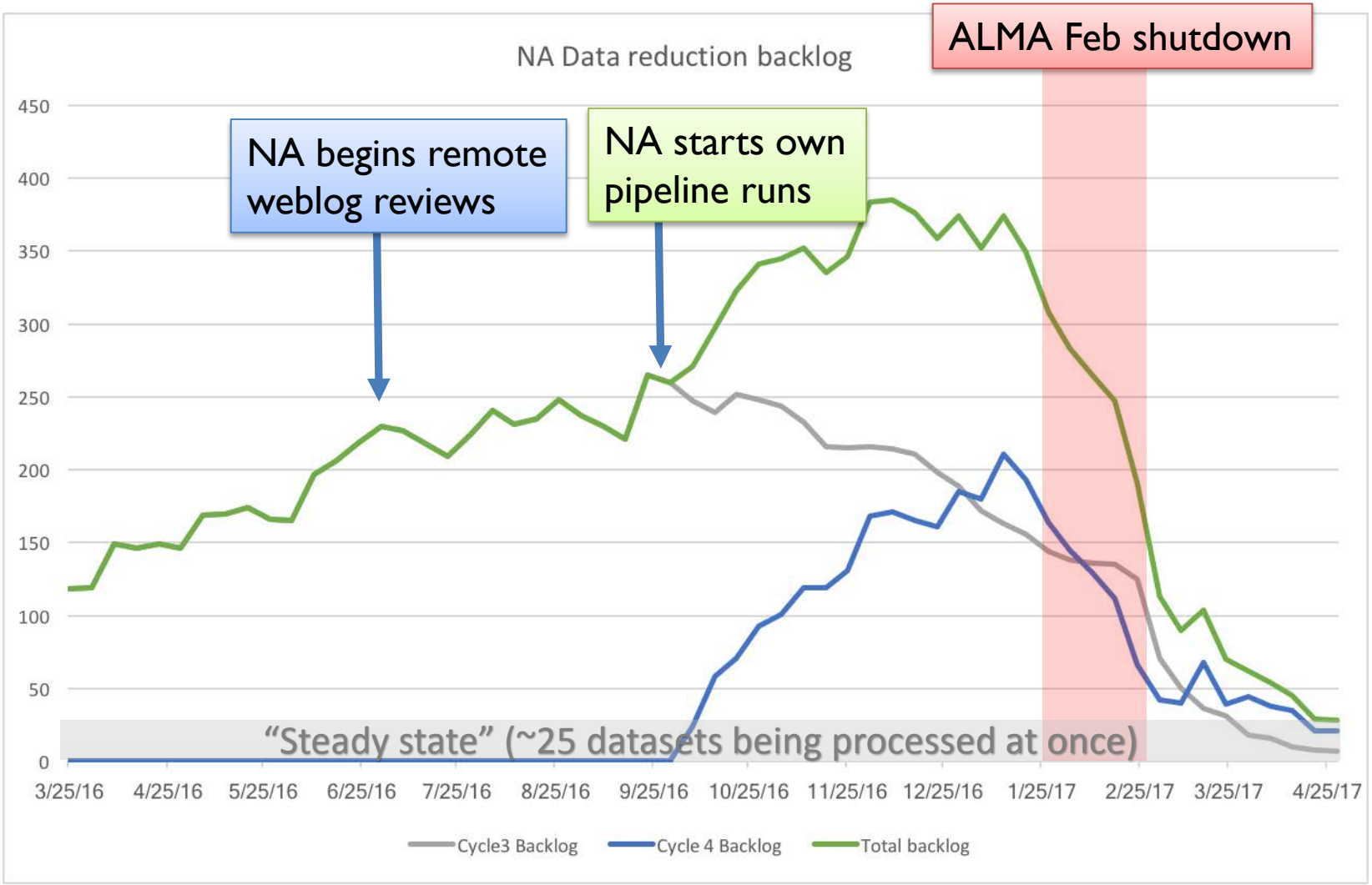
Wait time for Cycle 4 data (days). Current mean 48d.

Straightforward dataset timeline



Beating the backlog

- By the end of April we had hit steady state (data in/week~data out/week)



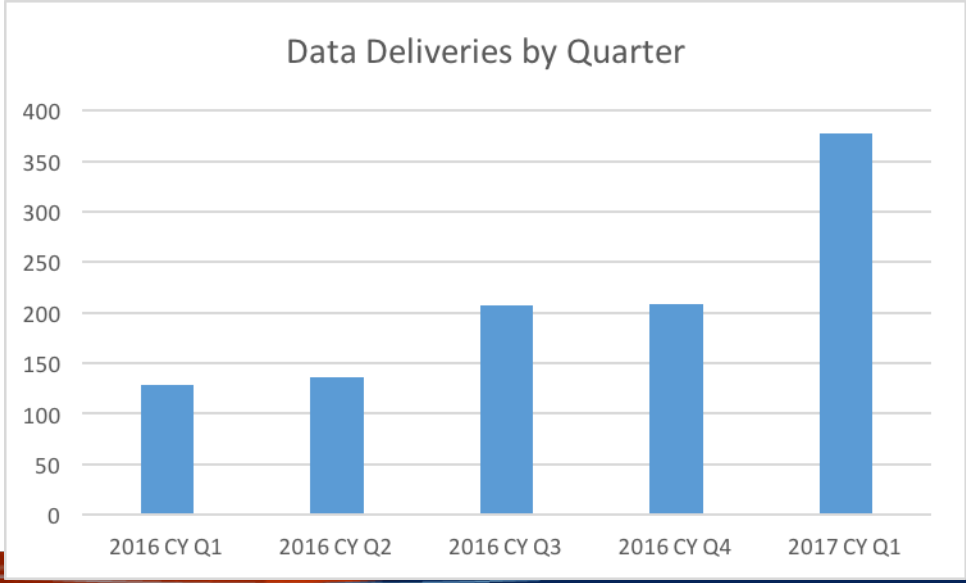
Data Reduction

The end of the backlog

- The advent of the imaging pipeline in Oct 2016 did not immediately see a reduction in the backlog – in fact it increased through December.
  - Learning curve for DAs to use the pipeline plus numerous problems with the Archive (NGAS) and Lustre filesystems.
  - By the mid December we had reached a steady state (helped by lower data rates from the telescope).
  - In February, the shutdown, plus some use of overtime for the DAs allowed us to fully assign all data for processing.
  - By the Cycle 5 deadline, only ~15 Cycle 3 and pre-shutdown Cycle 4 datasets remained to be processed out of the ~380 in the peak backlog.

Improvements In Efficiency

- Despite no net increase in the number of data reducers, the NAASC has improved the data delivery rate.
  - Switched to almost exclusively using DAs for data reduction, with scientists used for help with difficult datasets.
  - Once the most efficient way to run the pipeline was established, efficiency improved dramatically.



Data Reduction

Cycle 5

- Cycle 5: ~25% higher data rate, no more DA effort or computing nodes – need to automate more and make further efficiency improvements.
  - More batch processing to assign computing resources more efficiently. Working on memory estimates to optimize this.
  - Pipeline improvements should increase fraction of straightforward pipeline runs from ~50% to ~70%. Allows efficient running of the Cal+Img recipe to avoid breaks in the middle of the process.
  - Improvements to bookkeeping (40% of DA time). New software being deployed (AQUA/PT), though new system is actually less efficient in some respects (e.g. currently need to connect to JAO in real time), this is being worked.

Data Reduction
The ALMA Observatory as a whole

- NA's pipeline deployment benefited the Observatory as a whole, as we were able to pass on our experience operating the pipeline to the JAO and the other ARCs.
- Combined with more effort assigned to the data reduction at the JAO, this allowed the JAO to pipeline the EA and EU data, using the EA and EU ARCs for imaging difficult cases.
- EA and EU ARCs are also running the pipeline (though much less extensively).
- Together, these efforts have allowed the Observatory to reduce its backlog to ~250 datasets overall (compared to over 1000 in December).

Data Reduction Tiger Team

- A Tiger Team was convened at the end of February to address the issue of the data backlog, and put into place mechanisms to prevent a recurrence.
- Consisted of JAO management, subsystem scientists for archive, AQUA, scheduling and project tracker and data reduction managers at the ARCs.
- Outcomes:
  - Data reduction rate is limited mostly by the available human time to deal with manual calibrations and pipeline interventions (~2/3 of the available effort).
  - Thus ARCs will continue to help with data reduction for the next few cycles, until the pipeline is able to automatically process ~all the data.
  - Improvements to the process will be made to reduce the amount of bookkeeping.
  - Flagging will be enabled at the QA0 stage (immediately after observation) to identify and flag bad data before it enters the pipeline.

Summary

- ALMA data management remains an Observatory-wide challenge.
- Bulk of DA time is still spent on manual reduction or manual pipeline interventions.
- Are now finally hitting or exceeding our target to deliver pipelined Cycle 4 data within 30d, need to work on manual target of 45d.
- Currently in the midst of significant improvements to the pipeline and QA/tracking software.
- With these improvements we expect to be able to maintain our current ability to reduce the NA data as it is taken and avoid any further backlogs, even though the data rate in Cycle 5 and onwards will be higher.

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.