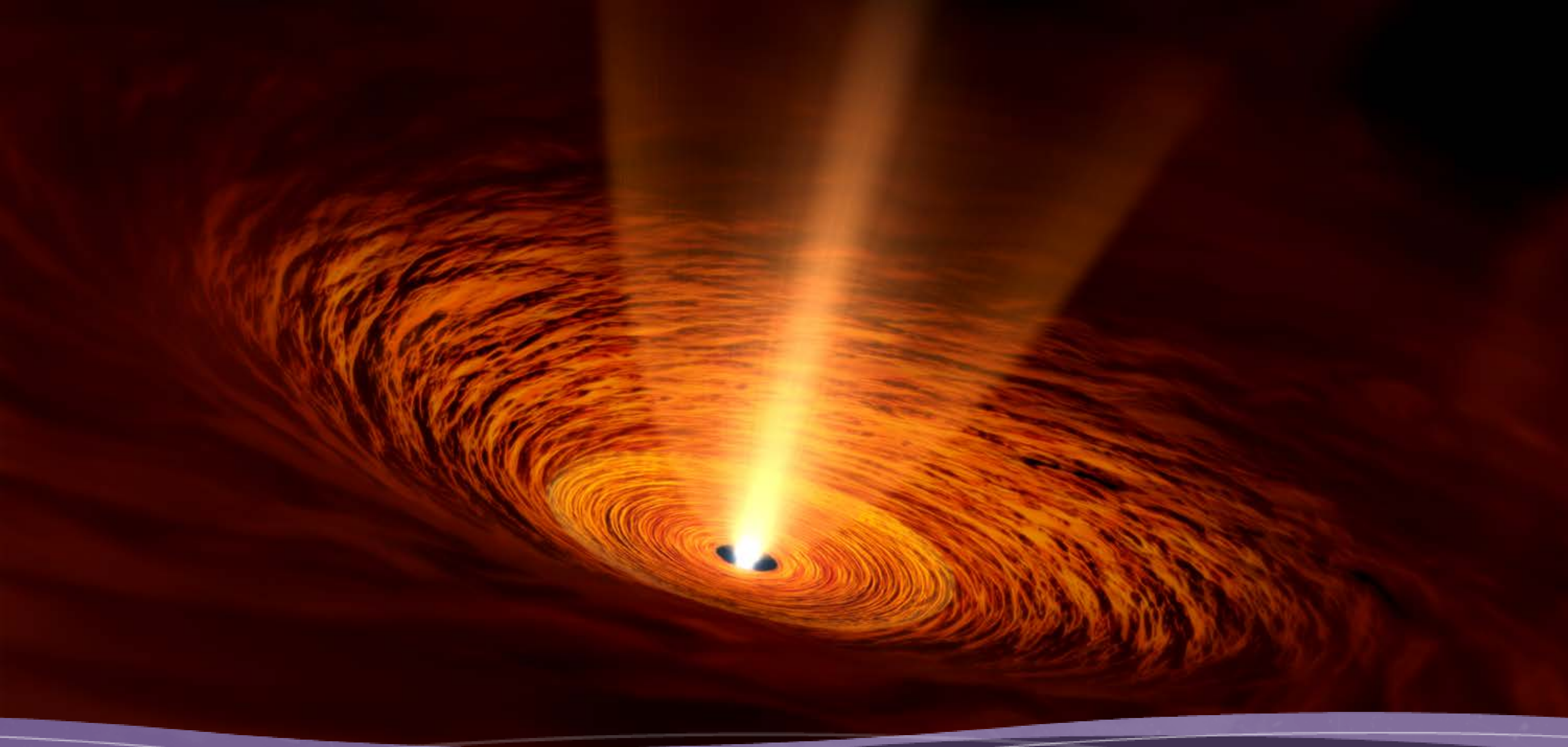


Data Processing and Workflow

Catarina Ubach & Mark Lacy

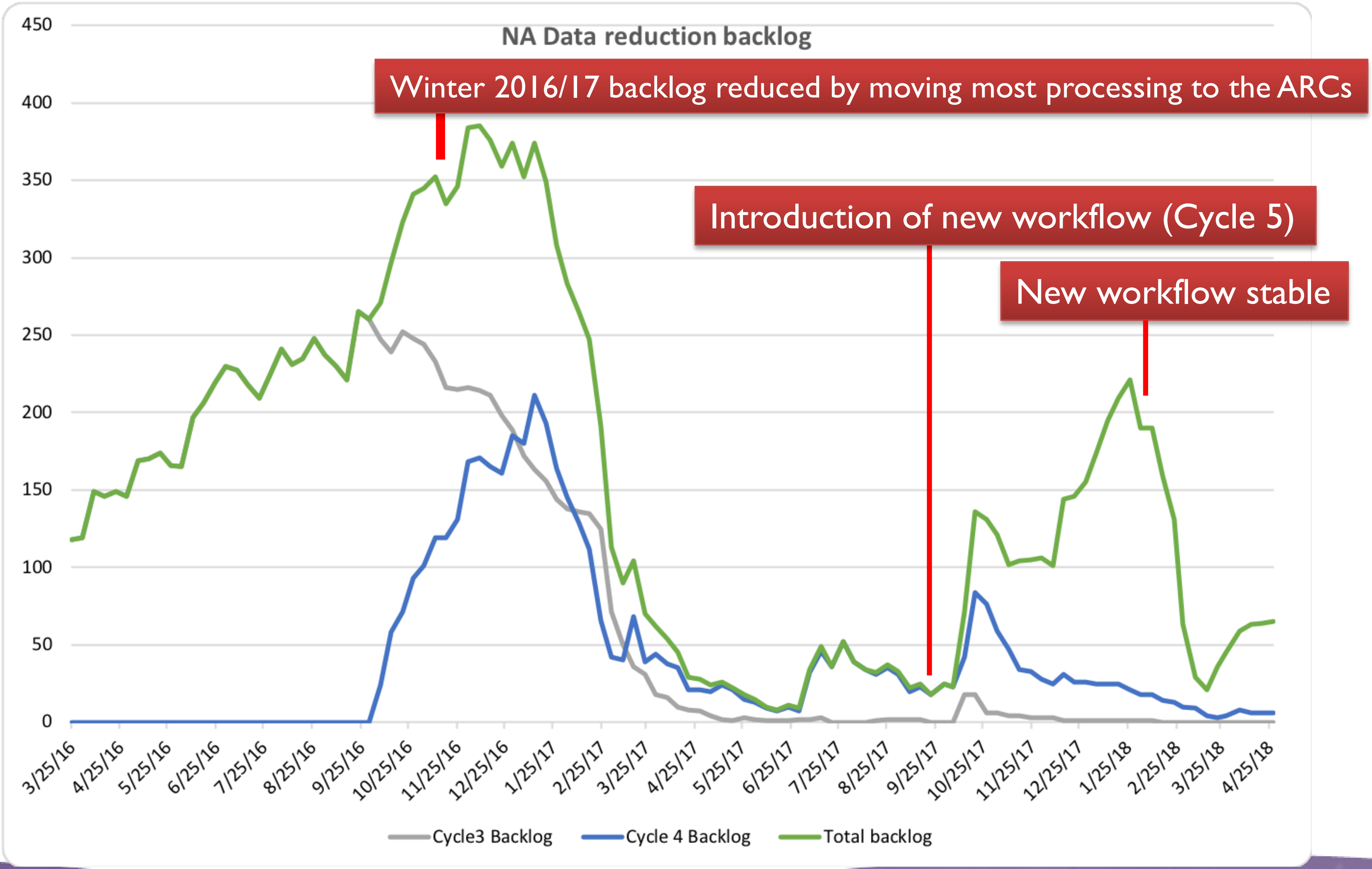
Overview

- New management for Data Management Group (DMG) at JAO (Theodorus Nakos).
- Data processing issues in Cycles 3 and 4 now largely overcome.
  - ARCs took on a large amount of pipeline processing, new short-term hires made.
- New data processing workflow introduced for Cycle 5 now uses the ALMA database as designed.
- Future data management challenges ahead as ARC effort ramps down over the next 1-3 years.

Major Initiatives/Recent Results

- New DMG lead Theodorus Nakos took over in October 2017, this transition was considerably assisted by NA personnel: John Hibbard as interim then deputy DMG lead from April 2017 until Dec 31st 2017, and Mark Lacy (from 1st Jan 2018) as C6 workflow scientist.
- A new processing workflow introduced in Cycle 5 that uses the ALMA database to track processing and automates many steps.
- DMG at JAO made significant purchase of hardware (disks and cluster nodes), and have established good relations with the Archive/Pipeline Operations Group, which has greatly improved their productivity.
  - An experiment of JAO running all pipeline data processing for 1 month (March 23rd-April 23rd) showed this hardware is sufficient, though ARC personnel were needed to supplement the current JAO staff for reviewing pipeline runs.
  - Now back to running pipelines at the ARCs (remote reviews are relatively inefficient).

User-Facing Issues – Data processing



Winter 2016/17 backlog reduced by moving most processing to the ARCs

Introduction of new workflow (Cycle 5)

New workflow stable

Response to UC 2017 – data delivery

- Major change in Cycle 4 was the full involvement of the ARCs in pipeline processing.
  - Limitations of hardware and personnel at JAO meant that throughput was not high enough to keep up with the data from the telescope.
  - JAO situation considerably improved in Cycle 5, but ARC effort still needed.
- Mean time to delivery for NA datasets is improving, current situation is stable (data out = data in).
  - Cycle 3 >100d
  - Cycle 4 63d
  - Cycle 5 33d (so far)
- Cycle 5
  - Problems with a new workflow provided a hiccup in processing efficiency from Oct until Feb, and another backlog, but it is now working fairly reliably (though it will need further improvement to make it robust in the long term).
    - Workflow uses the designed databases, rather than work-arounds, so meaningful states are now being set in the Project Tracker.

Response to UC 2017 – data product quality

- Problems with data reported by users are referred to as QA3s
  - Can be from the way the data were taken, reduced or packaged.
  - Over the past year we had a mean user-reported QA3 rate ~1% (9 cases vs 915 deliveries), though reports spiked just before the C6 proposal deadline.
  - A few problems are also caught after delivery, but before the user finds them (~2-3 this year).
  - Problems are mostly related to processing issues, roughly evenly split between pipeline and manual (probability of human error larger with manual reductions, but there are fewer of them).
  - Most user-reported problems (8/9) were resolved by reprocessing, or switching a pipeline processing to a manual one.
  - One case triggered an investigation that is still not clearly resolved.

Next 180 Days

- New software deployments at the end of May and at the start of Cycle 6:
  - Minor tweaks to the current workflow (e.g. automated delivery if only QA0-semipass data was taken for an SB, allowing a separate run of pipeline imaging.)
  - Expected to have little impact on data processing.
- Slightly more manual reductions expected in Cycle 6 (return of bandwidth switching), but should be low impact overall.
- Archive will start to serve individual products (FITS files etc) by October.
- "Raw data policy" in Cycle 6 – PIs can request raw data, but proprietary clock will start when they download the raw data, not when the products are made available.

Future

- The JAO is committed to producing Science Ready Data Products (SRDP) in the long term, a goal that is also consistent with NRAO's.
- In Cycle 6, we may see slightly longer than 30d to deliver data after it is fully observed, while diagnostics are being optimized for pipeline improvements (though raw data will be available for those who need rapid access).
- Actions we are taking in the medium term (Cycle 6-7) to improve the data processing workflow:
  - Making the QA2 process less burdensome
    - Improved weblogs: QA scores, better plotting
    - "Tiger team" to work with OT, scheduling, pipeline and DMG so beams & max. recoverable size as observed match the PI's request more closely.
    - Pipeline to adjust weighting robust parameter to more closely match the PI's request in Cycle 6.
  - Long term goals (Cycle 8+):
    - More automation of QA2 and data processing.
    - Application of machine learning/data science techniques to the weblog reviews and QA2 (starting with single dish).

Summary

- Data processing backlogs are now resolved, and we have a standard workflow that is both stable, and uses the production ALMA database rather than a work-around.
- New management at JAO has improved communications between the Data Management Group and Computing/Archive (APO), and significant computing infrastructure improvements have been made at JAO.
- Future challenges include improving automation of processing and the processing workflow to allow reassignments of NA Data Analysts to other NRAO priorities in the future.

NRAO

science.nrao.edu
public.nrao.edu
ngvla.nrao.edu

The National Radio Astronomy Observatory is a facility of the National Science Foundation operated under cooperative agreement by Associated Universities, Inc.