Lessons Learned from Developing and Operating the Kepler Science Pipeline and Building the TESS Science Pipeline

Jon M. Jenkins
NASA Ames Research Center

Monday September 4, 2017

WP11 Exoplanet Science Meeting
University of Warwick
Coventry UK
Overview

• What did it take to build the *Kepler* science pipeline?
• Major modifications to pipeline over lifetime
• High fidelity simulations
• Commissioning, commissioning, commissioning
• High performance computing
• Developing the TESS Science Pipeline
• Communication
• Summary
The Science Operations Center: What did it take?

- Design started in earnest in 2004 with launch in March 2009 and operations through May 2013 and reprocessing through 2017
- A total of ~100 person years of effort went into the first complete version of the pipeline (from pixels to planets)
- The staffing was at ~20 individuals per year through 2016, tapering off thereafter (~280 FTEs over project lifetime)
- Build 5.0 was the launch-ready software release
- There were 4 major builds thereafter, with substantive point releases to mitigate issues subsequently identified in flight or full volume re-processing
- Build 9.0, 9.1, 9.2, 9.3 really represented at least two full builds of effort (issues identified in full re-processing and in completeness and reliability processing)
- Unexpected instrumental effects/stellar variability/hardware failures motivated significant software modifications on orbit
A Search for Earth-size Planets

Kepler’s Science Pipeline

Artificial Transit & BEB Injection Machine
Raw Data

CAL
Pixel level Calibration

PA
Photometric Analysis

PDC
Pre-search Data Conditioning

TPS
Transiting Planet Search

Raw Light Curves & Centroids

Calibrated Pixels

Threshold Crossing Events (TCEs)

Corrected Light Curves

Diagnostic Metrics & Reports

Data Validation

Auto-Vetting
Applying machine learning to candidate evaluation

TCERT
Threshold Crossing Event Review Team + Robovetting

Planet, or dud?

>1,000,000 Lines of Code; 26 different Modules
Major Modifications

Every component of the science pipeline saw major evolution over mission

Pixel level calibrations:
- Updates based on actual electronics behavior
- Flagging of electronic image artifacts causing false positives

Identifying optimal apertures
- Use of reconstructed pointing
- Added ability to correct errors in Kepler Input catalog

Photometric analysis
- Major improvements to identifying cosmic rays

Pre-search Data Conditioning
- Development of Maximum a Posteriori approach
- Addition of multi-scale analysis
- Detection of Sudden Pixel Sensitivity Dropouts

Transiting Planet Search
- $\chi^2$ vetoes added

Data Validation
- Difference image analysis
- Ghost Diagnostic + other metrics
Short Timescale Instrumental Errors

Signature of a heater cycling on the reaction wheels 3/4

Kepler is sensitive to its thermal environment
Instrumental Effects in Photometry

A Search for Earth-size Planets

High pass filtered

Time JD - 2450000

Original
Corrected
We apply a Maximum A Posteriori approach as per Stumpe et al. 2014
End-End Model (ETEM) drove design of SOC and testing of entire ground segment

Simulated data were so good that we didn’t need to update the compression tables after launch (the achieved compression (~4.5-5 bits per pixel) was within 0.1 bits of ideal performance.)
Difference image analysis was key for Kepler for excluding false positives from background eclipsing binaries.

Especially important for bright, saturated (bleeding) targets

KIC 3542116

Fig. 5.— Direct images and difference images for KIC 3542116 during each of the three major transit features for each quarter in which they occur. Top panels show the mean calibrated pixel values in the aperture masks returned by Kepler for Q10, Q12 and Q13, from left to right. Bottom panels show the mean difference between the calibrated pixels in transit-feature-wide segments on either side of each transit, and the pixel values during each transit feature. The source of the transit features will exhibit positive values in these difference images. The colorbars indicate the pixel fluxes in units of $e^-/s$. Note that the pixels exhibiting the strongest positive deviations in the bottom panels occur primarily at the ends of the saturated and bleeding columns and are approximately clustered about the location of KIC 3542116, which is marked by a red star. KIC 3542117's location is marked by a red circle in each panel. These difference images indicate that the source of the transit features is co-located with KIC3542116 to within the resolution of the Kepler data.

We looked for stars located on the Kepler detectors which might be the source of any video cross-talk signals by inspecting the full frame images (FFIs) for the quarters during which the three most prominent transits occurred. There is a fairly bright, possibly saturated star near the edge of the optimal aperture on output 3 on the CCD on which 3542116 is imaged (it’s on output 1 in all cases), but the crosstalk coefficient is even smaller than for the other two outputs, $0.00001$, so that a 50% deep eclipse on this other star would be attenuated to a value of 5 ppm when its video ghost image is added to the direct image of KIC 3542116, assuming they are the same brightness. Furthermore, since the coefficient is negative, there would need to be a brightening event on the star on output 3 to cause a transit-like dip on output 1.

Fortunately, the largest crosstalk coefficient to the CCD output that 3542116 finds itself on is $+0.00029$, so given that the signal we are looking at is $\ll 0.1\%$, a contaminating star would need to be at least $10 \times$ brighter than 3542116 to cause a problem. If there were, it would be highly saturated and bleeding, which would make it difficult to square with the pixel-level analysis indicating that the source is associated with the pixels under 3542116, as the extent of the bleeding would be significantly larger than for KIC 3542116.

We also inspected the quality flags associated with the flux and pixel time series for KIC 3542116 and find that the situation is nominal with flags for occasional events such as cosmic rays and reaction wheel desaturations, but no flags for rolling band noise during the transit events.

Finally, we considered 'Sudden Pixel Sensitivity Dropouts' (SPSDs) in the data, which are due to radiation damage from cosmic ray hits on the CCD, as a possible explanation for the dips in flux that we observe. However, the shape and behavior of such dips do not resemble what we see (Thompson et al. 2016b). In particular, the SPSD events have drops that are essentially instantaneous, and therefore are much shorter than the $\ll 20$ and $8$ long-cadence points on the ingresses that we see in the deeper and more shallow dips, respectively. Moreover, the location of the SPSDs on the CCD chip would not plausibly align with the source location and its bleed tracks for each and every one of the dips. Thus, we also discarded this idea as well.

We take all these evaluations as strong evidence that the dips we see are of astrophysical origin and that KIC 3542116 is indeed the source of them. However, we cannot categorically rule out the possibility that the dips are caused by some unknown peculiar type of stellar variability in KIC 3542116 itself.

In spite of this caveat, we proceed under the assumption that the dips in flux are...
- Commissioning tools require special attention and data sets
- Effort for commissioning tools may be as great as that for major science pipeline modules
- Don’t leave commissioning tool development to the last
Pixel Response Function Characterization

**Kepler PRF**

A comparison of a synthetic image (left) with the actual flight image (right) for a short time interval for the Kepler flight image so the collection SNR will increase as pixels are added. After the background polynomial is subtracted from the pixel values, the target star is isolated.

**TESS PRF**

The same synthetic image is computed by including the pixel with the highest SNR. The result of this is a more optimal aperture that is due to the target star. The resulting collection SNR will increase as pixels are added. After the background polynomial is subtracted, the pixel collection with the highest SNR defines the optimal aperture. Fig. 3 shows an example of the dependence of the sky crowding metric on the number of bits in the analog-to-digital converter (= 14). The number of mesh lines is close to 36 × 36 = 1116 mesh lines, which is useful for identifying uncrowded stars.
Improving the Throughput

Some fast code; Some slow code

Step 1: Parallelize all code

Step 2: Make slow code fast(er)
64 hosts, 712 CPUs,
3.7 TB of RAM,
148 TB of raw disk storage
Hardware Architecture:
NAS Pleiades Supercomputer

5.34 Pflop/s peak cluster
211,872 cores
724 TB of memory
15 PB of storage
Characterizing completeness and reliability of software/people pipelines is extremely resource intensive. Kepler shipped the final light curve products in April 2015. We’ve spent the remainder of the time until present adding artificial transits, BEBs, scrambling the data temporally, inverting the light curves etc., etc.

Mapping completeness and reliability and characterizing the candidate vetting process is difficult.

Recommendation: Pursue machine learning for conducting or modeling the candidate vetting process.
Developing the TESS Pipeline

- ~13X pixel data rate over Kepler
- Leveraged heritage from Kepler pipeline
- Significantly lower cost (~46 FTEs over project lifetime)
- Significant speed improvements:
  - Colocated servers and storage with NAS Pleiades supercomputer
  - Moved pixel-level calibrations to C++
  - Sped up Presearch Data Conditioning by 10X
  - Originally projected 20+ days to process one sector
  - Complete pipeline requires ~5 days to process one sector
The TESS Project is distributed geographically with the Science Pipeline separated by a continent from the Science Office. Resolving data issues requires good communication between the Payload Operations Center, the Science Processing Operations Center and the Science Office.
New ideas for improving photometry/astrometry will emerge, both within the team and without

- “Halo” photometry on K2 data on Pleiades (White et al. 2017, MNRAS 471)

Preserving ability to re-process the pixel data with better algorithms and tuned parameters is a really good idea

Take advantage of the compressibility of your data

- *Kepler* achieved compression rates of 4.5 bits per pixel
- TESS should achieve compression rates of ~3 bits per pixel for 2 minute data and ~4 bits per pixel for 30 min FFIs
Summary

• Science pipelines require significant planning and effort
• Previous pipelines can be leveraged to reduce development time (but this does not reduce time required for V&V testing)
• Plan to rewrite the majority of the science code in light of unexpected in-flight characteristics/behavior/hardware changes
• High fidelity simulations are indispensable
• Determining $\eta_{\text{earth}}$ is computationally intensive and huge effort
• Give adequate attention to developing commissioning scenarios and associated tools
• Take advantage of data compression to increase the amount of pixel data downlinked from PLATO
SOC Cluster Architecture

6 Clusters:
4 Operations Clusters: Flight Ops, Quarterly, Monthly & Archive)
2 Test Clusters: LAB & TEST

Science Processing Pipelines

Long Cadence Photometry Pipeline

- DYN → CAL → TAD → PA → PDC → PAD → PMD → PAG

Transit Search Pipeline

- TPS → DV

FFI Pipeline

- CAL → PA

Short Cadence Pipeline

- CAL → PA → PDC