Processing and Managing the Kepler Mission's Treasure Trove of Stellar and Exoplanet Data

Jon M. Jenkins
NASA Ames Research Center
Moffett Field, CA USA
jon.m.jenkins@nasa.gov

***Abstract*—The *Kepler* telescope launched into orbit in March 2009, initiating NASA's first mission to discover Earth-size planets orbiting Sun-like stars. *Kepler* simultaneously collected data for ∼160,000 target stars at a time over its four-year mission, identifying over 4700 planet candidates, 2300 confirmed or validated planets, and over 2100 eclipsing binaries. While *Kepler* was designed to discover exoplanets, the long term, ultra-high photometric precision measurements it achieved made it a premier observational facility for stellar astrophysics, especially in the field of asteroseismology, and for variable stars, such as RR Lyraes. The *Kepler* Science Operations Center (SOC) was developed at NASA Ames Research Center to process the data acquired by *Kepler* from pixel-level calibrations all the way to identifying transiting planet signatures and subjecting them to a suite of diagnostic tests to establish or break confidence in their planetary nature. Detecting small, rocky planets transiting Sun-like stars presents a variety of daunting challenges, from achieving an unprecedented photometric precision of ∼20 parts per million (ppm) on 6.5-hour timescales, supporting the science operations, management, processing, and repeated reprocessing of the accumulating data stream. This paper describes how the design of the SOC meets these varied challenges, discusses the architecture of the SOC and how the SOC pipeline is operated and is run on the NAS Pleiades supercomputer, and summarizes the most important pipeline features addressing the multiple computational, image and signal processing challenges posed by *Kepler*.**

***Keywords*-data processing; high performance computing; software architecture; astronomy: extrasolar planets; astronomy: astrophysics;**

I. Introduction

The *Kepler* Mission was designed to discover Earth-size planets orbiting Sun-like stars through transit photometry: observing the small diminution of light that occurs when a planet crosses the face of its star from the observatory's point of view [1]. The amplitude of the planetary signal is minute, ∼100 ppm, and lasts from ∼1 hour to about half a day. This signature must be recognized against a variety of noise sources that are often much larger in amplitude, including instrumental effects such as shot noise and thermally-induced focus and pointing variations, and intrinsic stellar variability, including star spots and granulation noise. The transits repeat once per orbital period with an unknown phase, necessitating observations that can be carried out as long and as continuously as possible. In addition, most orbital configurations inhibit the observation of transits, as only a small fraction of possible inclination angles allow the planet to cross the face of the star from our point of view.

The *Kepler* Mission rose to these challenges with a 0.95-m aperture telescope that launched into orbit in March 2009 to conduct nearly continuous observations of up to 170,000 stars at a time in a single 116 square degree field of view (FOV) over a four-year mission. *Kepler* acquired data at 29.4-minute intervals for all target stars called long cadence (LC) targets and at 1-min intervals for up to 512 target stars at any given time called short cadence (SC) targets, and the resulting flux time series were typically >90% complete. The observations were organized into seventeen 93-day "quarters" by rotating the telescope by 90° every three months to keep the sunshade and the solar arrays properly oriented [2].[1] As of August 2016, 2679 planets have been discovered via transits, including 2330 planets discovered by *Kepler*.

The characteristics that made *Kepler* such a successful planet hunter also made it a near perfect stellar variability observation machine, providing important science results across a diverse set of stellar phenomena, including asteroseismology, gyrochronology, spot modulation, super flares, super novas, eclipsing binaries, "heartbeat stars," and relativistic boosting.

Undoubtedly, the success of *Kepler* was enabled by the exquisite instrument with its 95-megapixel camera and the benign orbit it occupies. Of equal importance is the science pipeline, which needed to keep up with the accumulating data volume, extract photometry at the 20-ppm level with a raw precision of ∼2%, and permit timely reprocessing of the data set as the pipeline evolved and its sensitivity improved.

The *Kepler* Science Operations Center (SOC) began development at NASA Ames Research Center over a 12-year period of time in 2004, continuing through the primary mission and well into the extended mission in light of three principle factors: 1) the stellar variability of the main-sequence stars in *Kepler*'s FOV proved to be twice as strong

[1] The first quarter, Q1, was only 34 days long due to the launch date and commissioning period. The last quarter, Q17, was only 31 days long, due to the mission-ending loss of reaction wheel #4, and contained a 10-day rest period to attempt to increase the lifetime of this reaction wheel.

as expected, based on long-term observations of the Sun [3], [4], 2) instrumental effects caused both by radiation damage and by electronic image artifacts triggered an overabundance of false alarms and threatened to overwhelm the system [5], and 3) the interplay of the intrinsic stellar signatures and instrumental signatures required the development of more sophisticated approaches to identifying and removing systematic errors than were available in the original pre-launch pipeline design [6].

These issues stimulated significant research and development of new algorithmic approaches for virtually every module of the science pipeline. As the pipeline evolved, the data needed to be reprocessed, and this, too, was a challenge.While the the 700-node computer cluster used for processing the *Kepler* data was able to keep up with the data as it was downlinked every month, it could not reprocess the accumulating data record in a reasonable amount of time. This motivated the SOC to develop new software infrastructure in order to be able to routinely process and reprocess data on the NASA Advanced Supercomputing (NAS) Division's Pleiades supercomputer.[2]

This paper is organized as follows: Section II presents some of the most compelling examples of astrophysics achieved through the *Kepler* Mission. The high-level architecture of the SOC is described and discussed in detail in Sec. III. How the *Kepler* data are processed on the NAS Pleiades supercomputer is described in Section IV. Conclusions are presented in Section V.

II. Astrophysics with *Kepler*

The impact of *Kepler* on exoplanets can perhaps only be eclipsed by its impact on astrophysics in general, as measured by the relative number of exoplanet publications to astrophysics publications based on *Kepler* data. In this section we provide a few examples of how *Kepler* has contributed to the study of stellar phenomena.

Asteroseismology *Kepler* revolutionized the field of asteroseismology of solar-like stars by permitting the p-mode (or pressure mode) oscillations of these stars to be observed for >500 stars by using SC observations, whereas prior results using Doppler techniques were limited to ∼20 stars [8]. Asteroseismology can provide estimates of stellar mass and radius to within a few percent for sufficiently bright stars, and also places good constraints on stellar age, thereby significantly enhancing the precision of the planetary parameters for any transiting planets discovered orbiting such stars. While p-mode oscillations of solar-type stars exhibit typical periods of several minutes, red giants oscillate at much longer periods and are observable with *Kepler*'s LC data; [9] documents results for over 13,000 stars.

Fig. 1 shows a Hertzsprung-Russell diagram for 15,000 stars exhibiting p-mode oscillations observed by *Kepler* [7].

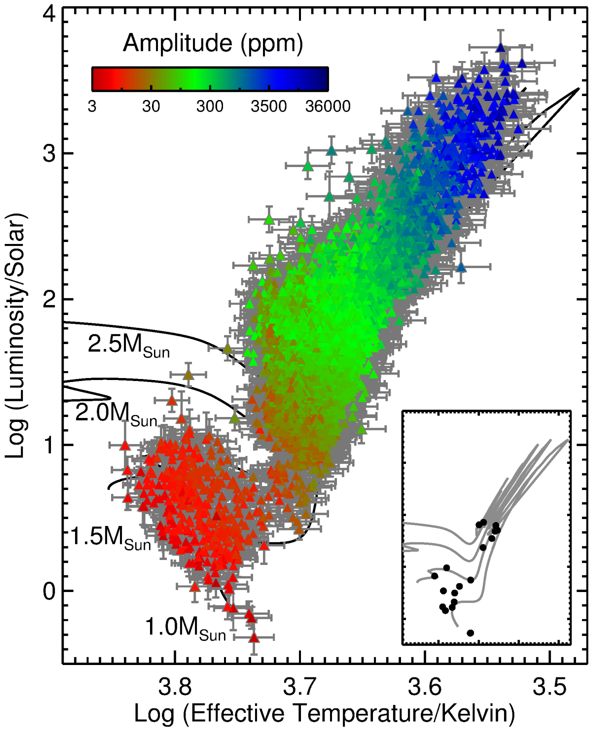

Figure 1: A Hertzsprung–Russell diagram for 15,000 stars exhibiting p-mode oscillations observed by *Kepler* displaying log luminosity vs. log effective temperature. The points are colored by the amplitudes of the stellar oscillations, which vary from 3 ppm to ∼3600 ppm. These results illustrate the fact that the amplitudes vary with the mass and size of the star. The inset shows similar results for ∼20 stars obtained prior to 2008. From Fig. 3 in [7].

Extended studies can probe internal differential rotation structure of sub-giants and red giants, and distinguish between H shell-burning red giants from He-burning ones [10].

RR Lyrae Stars Another exquisite science result enabled by *Kepler* is represented by observations of RR Lyrae stars, including the eponymous RR Lyr itself [11]. RR Lyrae variable stars are low density, He-burning stars, and are considered standard candles as there is a good relationship between their pulsation period and their intrinsic brightness in the infrared (though not at visual wavelengths). Some RR Lyrae stars exhibits the Blazhko effect, whereby the amplitude of the oscillations experiences periodic modulations over timescales much longer than the pulsation period. An example of this is given in Fig. 2a. RR Lyrae periods are ∼0.5 days, making them difficult to study in detail from ground-based observations due to the day/night cycle.

Heartbeat Stars *Kepler* serendipitously discovered a new class of binary star system in highly eccentric (non-circular) orbits. KOI-54 was the first example and is composed of two nearly identical A-type stars in a nearly face-on 41.8-

[2] Pleiades currently has 227,808 computer cores and 828 TiB of memory (http://www.nas.nasa.gov/hecc/resources/pleiades.html).

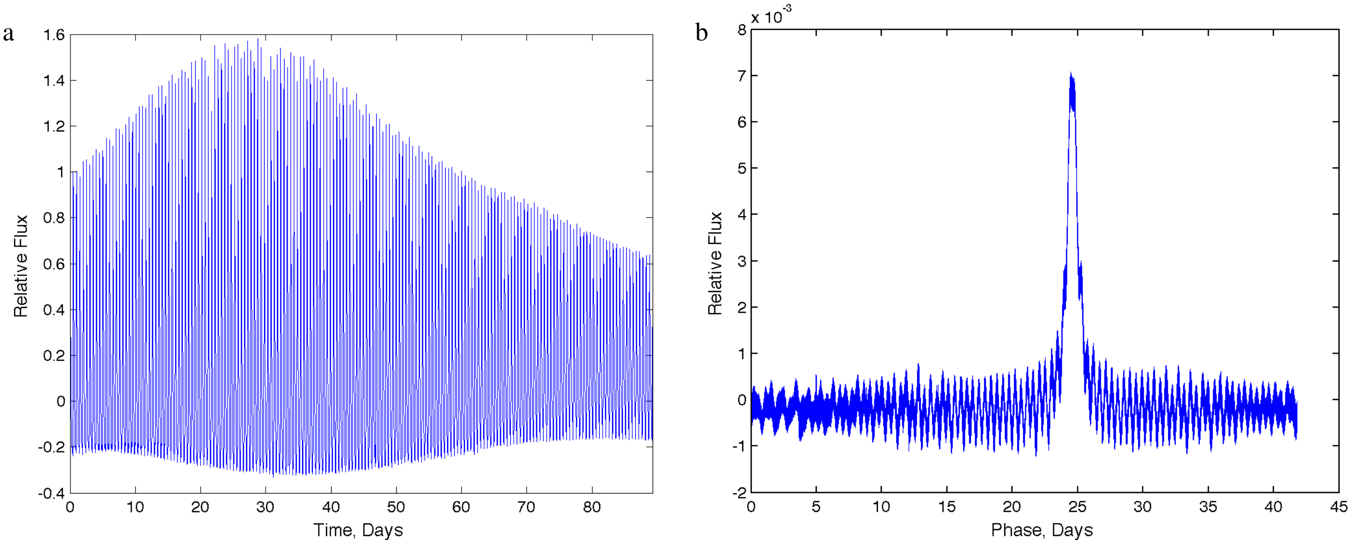

Figure 2: Light curves for two oscillating stars observed by *Kepler*. a. Observations of the star KIC 7671081, an RR Lyr star exhibiting amplitude modulation synonymous with the Blazhko effect. b. Phase-folded light curve for the binary system KOI-54, which is in a highly eccentric (non-circular) orbit.

day period orbit with an eccentricity $e = 0.83$ [12]. The tidal distortions driven by their close passage to one another drive oscillations, with the 91st and 92nd harmonics of the orbital frequency being most dominant, providing a 7X increase in brightness of the system as the two stars make their closest approach to one another (see Fig. 2b).

Star Spin Down Rates – Gyrochronology Cool stars lose angular momentum and spin down with time with the spin down rates depending chiefly on age and mass. *Kepler*'s nearly uninterrupted photometric measurements of unprecedented duration (years), temporal resolution (minutes), and precision (ppm) enabled the *Kepler* Cluster Study to measure the spin down rate of cool, main-sequence stars via observations of the ∼1 Gyr-old star clusters NGC 6811 [13] (see Fig. 3) and the ∼2.5 Gyr-old cluster NGC 6819 [14]. These studies indicate that the color-period diagrams for these clusters display a single tight rotational sequence from mid-F to early-K spectral type, extending earlier results for younger clusters up to 2.5 Gyr suggesting that cool stars populate a thin surface in rotation–age–mass space. This indicates that ages can be estimated with a precision of ∼10% for large numbers of cool Galactic field stars.

Many of these astrophysics results depend on the *Kepler* science pipeline's ability to identify and remove instrumental signatures while retaining intrinsic astrophysical signals. Major effort was expended to develop a Bayesian approach to this problem, leading to the Presearch Data Conditioning (PDC) *Maximum A Posteriori* (PDC-MAP) module, discussed in Section III-C4.

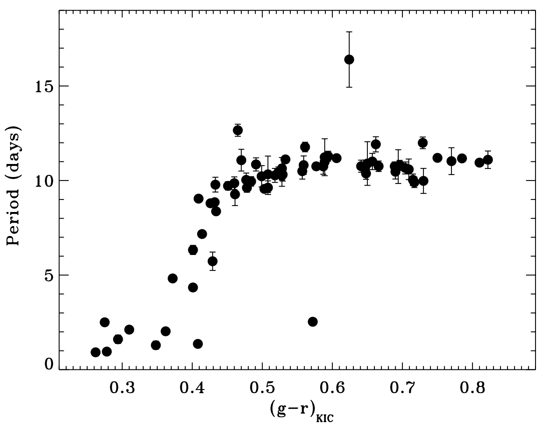

Figure 3: Stellar rotation periods measured for 71 dwarf stars in the open cluster NGC 6811 as a function of color. From Fig. 4a in [13].

III. The Science Operations Center

The SOC consists of several elements: 1) a pipeline infrastructure coded in Java that ingests the science data, controls the science processing, and writes the archive data products to files in the archive file format, 2) the science pipeline itself, 3) a target management system that contains a catalog of target and field stars and their characteristics, and that identifies the pixels of interest for each target star and associated on-chip collateral data, and 4) a suite of commissioning tools used to obtain or validate various calibration models and pre-flight instrument characterizations. Fig. 4 shows the high-level architecture of the SOC.

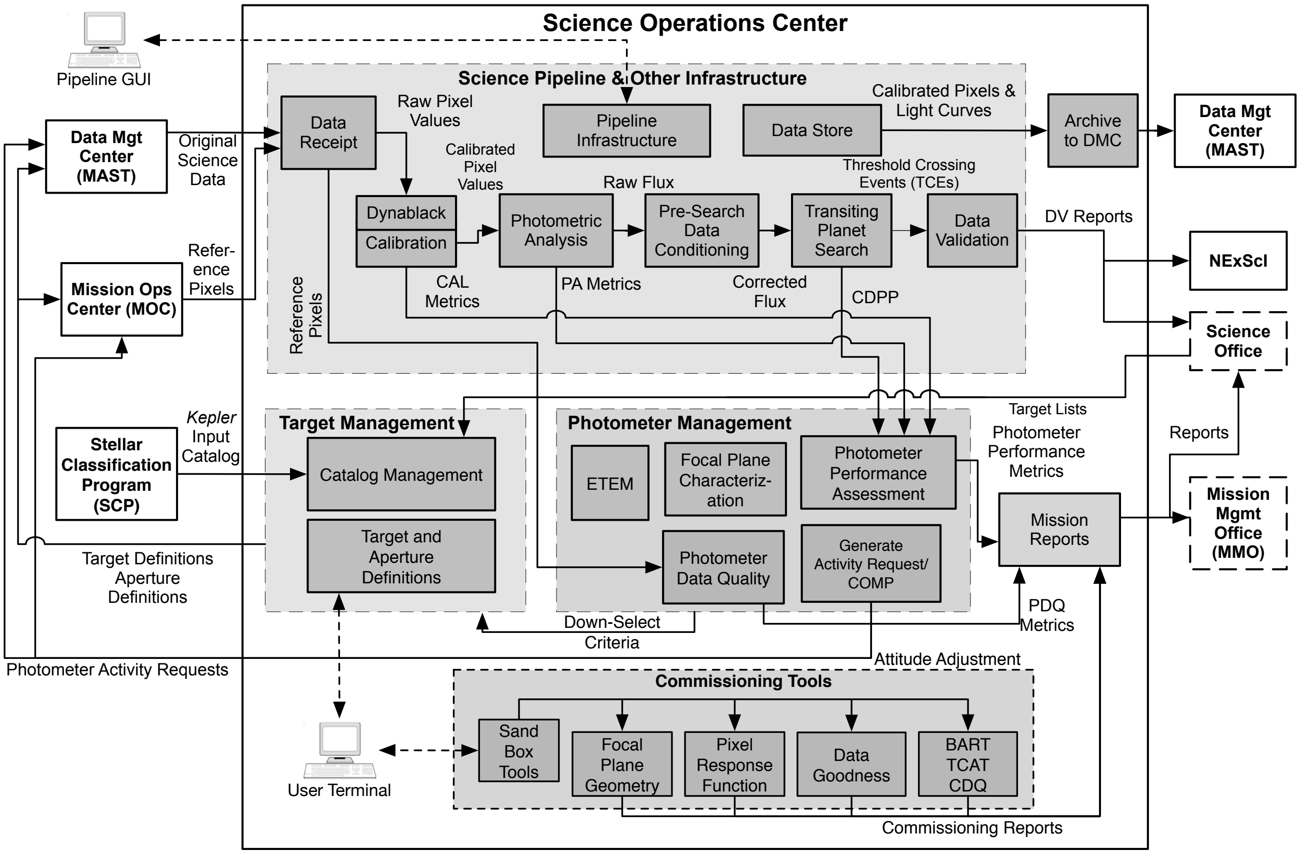

Figure 4: Architecture diagram of the *Kepler* Science Operations Center indicating the 23 major components comprising the science pipeline, target management, commissioning tools, and photometer management functions.

A. Software Infrastructure

1) Pipeline Infrastructure (PI): PI provides fully automated distributed processing of science data and sequencing of modules based on the results of previous modules [15], [16]. Features include a customizable unit-of-work that controls how the data are distributed across the cluster, a configuration management and versioning system for algorithm parameters and pipeline configurations, and a graphical user interface for the configuration, execution, and monitoring of pipeline jobs. PI provides scalability for running the pipeline on a developer workstation, a large cluster of computing nodes, as well as on the NAS Pleiades supercomputer.

2) Data Receipt (DR): DR provides data ingestion and automated pipeline launch capabilities, and is divided into two main components: 1) a generic layer that watches for new files, dispatches the proper handler, and launches pipelines and 2) a plug-in layer for specific data types.

3) Data Store (DS): The *Kepler* DS is a transactional database management system for arrays, sparse arrays, and binary data [17]. DS consists of a custom array data base (ADB) and an Oracle database. The data volume of 4-year *Kepler* data set is ∼32 TiB.

4) Archive (AR): AR generates the files in the archival format archived to MAST and made available to the science team and greater astronomical community. The *Kepler* archival products include calibrated pixels, simple aperture photometry, systematic-error-corrected photometry, astrometry (centroids), and associated uncertainty estimates. Target pixel files contain the pixel data, both original and calibrated, for each target organized as image data. These files also include information about sky background flux and cosmic ray hits detected by the pipeline. The transit-like features identified by the transit search and the results of DV's diagnostic tests are archived as XML files along with PDF reports to the Exoplanet Archive managed by NASA's Exoplanet Science Institute. These PDF files contain a wealth of information regarding each transit-like signature.

5) Mission Reports (MR): MR provides a web-based interface to a library of reports concerning the pipeline and its processes that can be generated on the fly.

6) Target Management: The *Kepler* target management functionality consists of two components:

1) Catalog Management (CM) contains the *Kepler* Input Catalog (KIC) provided by the Stellar Classification Program

[18] and subsequent updates [19]. CM contains the characteristics of the target stars, and background field stars, such as location (right ascension, declination), effective temperature, surface gravity, radius, mass, and proper motion; and

2) Target and Aperture Definitions (TAD), which formulates the target definitions specifying which pixels need to be stored and downlinked by the *Kepler* spacecraft. Tad also formulates the 1024 mask definitions used to capture the pixels of interest of each target. The associated sub-module, Compute Optimal Apertures (COA), predicts the pixels of interest for extracting photometric measurements from the CCD images for each target star in the SOC pipeline [20]. *Kepler* had very tight margins for the pixel data stored onboard and returned to the ground: only ∼6% of the pixel data could be stored onboard for later downlink.

B. Photometer Management

This suite of software contains the calibration models used by the science pipeline as well as modules that monitor the performance of the photometer.

1) Photometer Performance Assessment (PPA): PPA assesses the health and performance of the instrument based on the science data sets collected each month, identifying out-of-bounds conditions and generating alerts [21]. The metrics are tracked and trended and reported numerically as well as in a PDF report, and include photometric precision, brightness, black level, background flux, smear level, dark current, cosmic ray counts, outlier counts, centroids, reconstructed attitude, and the difference between the reconstructed and nominal attitudes. These metrics are tracked and trended by PPA and the numerical results are persisted to the DS as well as populating a PDF report. PPA results are used to identify and set data anomaly flags required for archival processing.

2) Photometer Data Quality (PDQ): PDQ provides a "quick look" assessment of the health and performance of the instrument through data downlinked at X-band twice-weekly [22]. PDQ also assesses the validity of the spacecraft pointing after each return to science attitude and issues a corrective "tweak" if any target star is more than $0.4''$ from its desired location.

3) End-To-End Model (ETEM): ETEM is a suite of software that generates synthetic flight-like data for *Kepler* with a high degree of fidelity, including matching the formats of the science data at each ground segment interface, from the solid state recorder (SSR) onboard the spacecraft, through the MOC and the DMC [23], [24]. ETEM was indispensable in testing the entire *Kepler* ground segment as well as for designing, implementing, and testing the SOC. ETEM simulates the astrophysics of planetary transits, stellar variability, background and foreground eclipsing binaries, cosmic rays, and other phenomena.

4) Focal Plane Characterization (FC): FC consists of a set of database tables, persistence classes, and associated handling code that manages the calibration models used to process data and manage target definitions [25].

5) Compression (COMP): Pixel data compression tables are generated by two components named the Huffman Generator (HGN) and the Huffman Aggregator (HAG) modules. The data compression scheme involves three steps: 1) re-quantizing the data so that the quantization noise is approximately a fixed fraction of the intrinsic measurement uncertainty (which is dominated by shot noise for bright pixels), 2) taking the difference between each re-quantized pixel value and a baseline value that was updated once per day, and 3) entropic encoding via a length-limited Huffman table [26]. Typical compression rates of 4.5–5 bits per pixel measurement were achieved throughout the *Kepler* Mission, allowing for >66 days of data to be stored on the SSR and decreasing the time required for DSN contacts.

C. Science Pipeline

The science pipeline calibrates the original data from *Kepler* and produces the archival data products. It conducts the transit search and constructs diagnostics used to prioritize and rank the planetary candidates for follow-up observations.

1) Calibration (CAL): This module operates on original spacecraft pixel data to remove on-chip artifacts such as smear from the shutterless readout [27]. Traditional CCD data reduction is performed (removal of instrument/detector effects such as bias and dark current and flat field), in addition to pixel-level calibrations. CAL operates on SC and LC data as well as Full Frame Images (FFIs – nominally acquired once per month), and produces calibrated pixel flux time series, associated uncertainties, and metrics that are used in subsequent pipeline modules.

2) Target and Aperture Definitions (TAD): In the context of the science pipeline, TAD updates the photometric apertures based on the reconstructed pointing history obtained by centroiding a fiducial set of bright, unsaturated targets on each 29.4 min cadence.

3) Photometric Analysis (PA): PA measures the brightness of the image of each target star on each frame. It also fits and removes background flux due to zodiacal light and the diffuse stellar background, identifies and removes cosmic rays all target star apertures, and measures the photocenter or centroid of each target star on each frame. PA also uses PRF-fitting to measure precisely the location of ∼200 bright, unsaturated target stars on each CCD readout area in order to establish the pointing and focus of each camera. This information is used to update the photometric apertures [28].

4) Presearch Data Conditioning (PDC): PDC performs a critical set of corrections to the light curves produced by PA, including the identification and removal of instrumental signatures caused by changes in focus or pointing, and step discontinuities that result occasionally from radiation events in the CCD detectors. PDC also identifies and removes isolated outliers and corrects the flux time series for crowding effects and for the fact that not all the light from a star can be captured by a finite aperture [29], [30]. PDC employs a multi-scale MAP approach to identify and remove systematic errors, allowing it to retain important astrophysical signals in the face of much larger instrumental effects, as illustrated by Fig. 5.

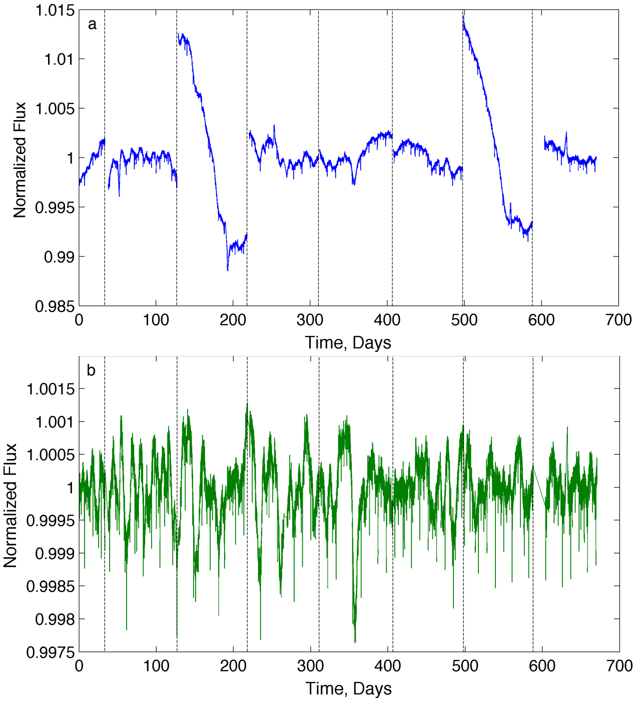

Figure 5: a. The original simple aperture photometry light curve for Kepler-20, a star hosting five transiting planets, including the first Earth-size planet discovered by *Kepler* [31]. b. The systematic error-corrected light curve produced by PDC using a MAP-based approach, showing good preservation of the transit signatures and rotational modulation of the star's brightness by star spots.

5) Transiting Planet Search (TPS): TPS implements a wavelet-based, adaptive matched filter algorithm to detect signatures of transiting planets [32], [33]. TPS stitches the ∼93-day light curves together for stars observed on consecutive sectors prior to searching for planets. TPS also provides estimates of combined differential photometric precision (CDPP), a key performance diagnostic for transit survey missions, on timescales of transits [34].

6) Data Validation (DV): This component performs a suite of diagnostic tests on each transiting planet signature identified by TPS to make or break confidence in its planetary nature. These include a comparison of the depth of the even transits to the odd transits, an examination of the correlation of changes in the photocenter (centroid) of the target star to the photometric transit signature, a statistical bootstrap to assess confidence in the detection, difference image centroiding to rule out background sources of confusion, and a ghost diagnostic test to rule out optical ghosts of bright eclipsing binaries as the source of the transit-like features. These tests can determine if the transit signature is likely to be due to a background eclipsing binary whose diluted eclipses are masquerading as transits of a planetary body. DV also calls TPS to search the residual light curve for evidence of additional transiting bodies after fitting and removing the first planetary transit signature from the light curve. This process is repeated until TPS fails to identify another transit signature.

D. *Commissioning Tools*

Several tools were developed and deployed specifically for the commissioning phase of the *Kepler* Mission.

1) Focal Plane Geometry (FPG) and Pixel Response Function (PRF): FPG and PRF were used to determine the detailed sky to pixel mapping coefficients and the shape of the PSF, respectively, across each of the 84 individual CCD readout channels. The FPG coefficients include terms for pincushion distortion. PRF constructed five individual PRFs for each CCD readout area in order to capture nonuniformity in the focus and PSF. PRF models at intermediate locations are obtained by interpolation. The pipeline uses these model waveforms to monitor the locations of the brightest, unsaturated 200 stars on each channel to reconstruct pointing and capture distortion due to focus changes.

2) Pixel Overlay On FFIs (POOF): This tool allows the user to retrieve *Kepler* FFIs and overlay the aperture masks from target tables on the images, along with information about the stellar targets themselves. This enabled the validation of target tables early in the *Kepler* Mission.

3) Data Goodness (DG): The DG tool allows the user to verify the quality of FFIs.

4) BART, TCAT and CDQ: These tools were designed specifically to confirm that the behavior of certain electronic artifacts discovered pre-launch remained consistent after launch.

5) Sandbox Tools (SBT): SBT allow *Kepler* personnel to make queries against the DS on their own workstations.

E. *Hardware*

This section describes the hardware used to support the pipeline.

1) Cluster Worker Machine: Cluster worker machines are used to serialize data inputs for MATLAB executables that run on Pleiades and to parse the outputs generated from those pipeline modules. Worker machines can run these MATLAB executables locally, but at a much smaller scale. This is done for less numerically intensive modules such as PPA and AR.

Cluster worker machines can be moved between different clusters in order to provide for some redundancy. We procured worker machines with 24 cores (48 including hyper threads) and 768 GiB of RAM. Temporary task file storage is on a local NFS. This means local storage on the worker machine is not a bottleneck for either storage capacity, performance or availability. Two 10 GiB Ethernet network interfaces are present on the worker machines.

F. Cluster Datastore Machine

Each cluster also has a dedicated datastore machine that are similar to the cluster worker machines with the addition of two 8 GiB Fibre Channel host bus adapters. Oracle and ADB share this machine. The *Kepler* Storage Area Network (SAN) has a storage capacity of ∼200 TiB and a theoretical transfer rate limit of 2 GiB $\sec^{-1}$, which is sufficient to copy the entire contents of the storage array in about 10 minutes. Practically, other parts of the architecture are the limiting factor. For *Kepler*, the SAN is often not the bottleneck as the number of concurrent I/O operations is limited by the number of disks rather than by the network.

ADB is allocated 64 GiB of RAM with Oracle using the remainder. ADB uses most of its memory to cache b-tree indices and for temporary buffers. Index blocks are used to locate array blocks on disk; this scales with the number of independent arrays in working memory which is typically largest during processing for CAL. The total memory required for *Kepler* is ∼32 GiB but 64 GiB are available. Oracle tends to be bottle necked on disk I/O rather than processor power.

1) Data Storage: Data storage is handled via a SAN. This is a dedicated network for the transmission of blocks of data to and from datastore machines. We use a storage array with ∼200 7.2k RPM hard disks. The storage array presents the view of one or more virtual block devices to each host known as volumes or logical unit number (LUN). Each volume is in fact a combination of disks placed in a RAID 6+0 configuration. This allows for each LUN to be striped across all the drives in the array so that each can access the full number of I/O operations. The storage array can provide approximately 15K I/O operations per second. A volume can also be snapshot, which is a point-in-time copy of a base volume. Modifications to snapshots have a copy-on-write feature which means space is only allocated for modifications. Failed disks can be replaced with reserve space on the remaining working disks. At a minimum, two disks can fail without a loss of data. In practice, many more disks have failed without data loss.

IV. Running the SOC Pipeline

The SOC science processing pipeline is actually configured as multiple pipeline segments based on the dataset types that they process and the frequency at which they run, as indicated in the typical pipelines depicted in Fig. 6. Each pipeline is a directed graph of pipeline modules.

Organizing these processing steps as separate pipelines provides flexibility without complicating the pipeline configuration. This flexibility also allows other scenarios to be implemented. The set of pipeline segments and the modules which they contain are completely configurable using the pipeline GUI. The set of available modules (the module library) is also configurable. This architecture enables modules to be easily updated, added, or removed without code changes (other than to the modules themselves).

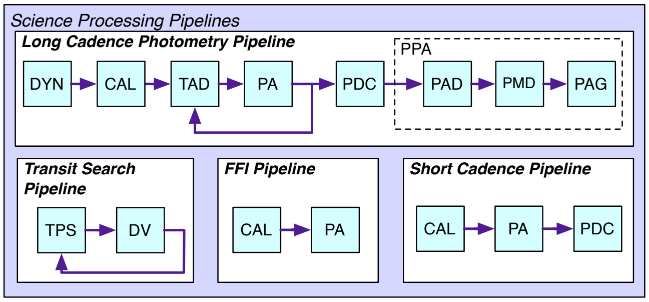

Figure 6: There are four major *Kepler* pipelines. Top panel: The photometry pipeline for LC data indicating that TAD is run twice, once before the pointing is reconstructed by PA, and once afterwards. Only LC data are processed by PPA for attitude determination (PAD), for metic determination (PMD) and for aggregation of the metrics (PAG) and generation of the reports. Left bottom panel: The transit search pipeline for the LC data, indicating that DV can call TPS multiple times. Middle bottom panel: The FFI pipeline. Right bottom panel: the SC data pipeline.

A. Unit of Work

Since the pipeline algorithms can be computationally intensive, and the large data volumes involved, the pipeline can be run on multiple machines. To this end, the pipeline can run on a cluster of local worker machines or a set of remote machines on the NAS Pleiades supercomputer. When a new pipeline is launched, the work is divided into units that can then be distributed to individual worker machines. This unit-of-work (UOW) can be configured to bin the input data by cadence, CCD output, and/or target. The design goal is to use the UOW as a tuning knob to maximize concurrency. This knob is adjusted based on how many machines are available and how long a UOW takes to process.

For execution on Pleiades, a UOW is further broken into subtasks so that work can be distributed across the tens of thousands of nodes and cores in Pleiades. Additional tuning parameters control how many subtasks can execute on each node so as to best take advantage of the memory and cores present in each node. Table I lists units of work and subtasks for several pipeline modules.

B. Coordination of Work

1) Local Cluster: When a new pipeline is launched, a pipeline task is created for each UOW for each module in the pipeline. Pipeline tasks are scheduled for asynchronous execution using a distributed message queue also known

Table I: Units of work to the subtask level for various pipeline modules.

Pipeline Module	**Binned by**
CAL	cadence interval, CCD output, CCD row(s)
PA	month (SC) or quarter (LC), CCD output, individual targets
PDC	month (SC) or quarter (LC), CCD output
TPS	individual targets
DV	individual targets

as Message Oriented Middleware (MOM). At the start of execution, a message is placed on the queue for each pipeline task. Once the messages are on the MOM queue, the next available worker machine will pull the next message off the MOM queue and execute its task. Any worker is able to execute any pipeline task because each worker machine has access to all of the science modules and the pipeline infrastructure services. This design allows worker machines to be easily added for increased processing throughput.

C. Remote Execution

Remote execution is distinguished by the use of third-party authentication and connection tools in order to execute pipeline tasks on Pleiades. In this case the pipeline worker processes remain local and files are transmitted over secure shell (ssh) via the Multi Mission Operations Center (MMOC) network. A remote queuing system (RMOM), allocates super computer nodes to subtasks. Pipeline modules can generate a dependency graph that expresses the dependencies between subtasks, such as the fact that the image motion information needs to be generated by PA prior to assigning the final photometric apertures via TAD. The RMOM obeys this dependency graph and so as many independent subtasks are run on at least as many available processing nodes (Fig. 7). There are additional parameters that determine the number of concurrent subtasks that can execute on a processing node. This is usually limited by the memory-to-core ratio of the type of subtask being executed. While not limited to coordinating MATLAB processes, these are the types of processes that are executed on the Pleiades.

D. Triggers

New pipeline instances are launched using pipeline triggers, which associate pipeline parameters with specific pipeline instances. These triggers are part of the pipeline configuration and are created by the pipeline operator using the pipeline GUI. Triggers can also be used to launch pipelines manually, as in the case of reprocessing, or automatically on a particular schedule or when the input data become available. These data-available triggers allow the various pipeline types to be chained together so that complete, end-to-end science processing can be automated. For example, the photometry pipeline can be configured to run when new data are delivered from the spacecraft.

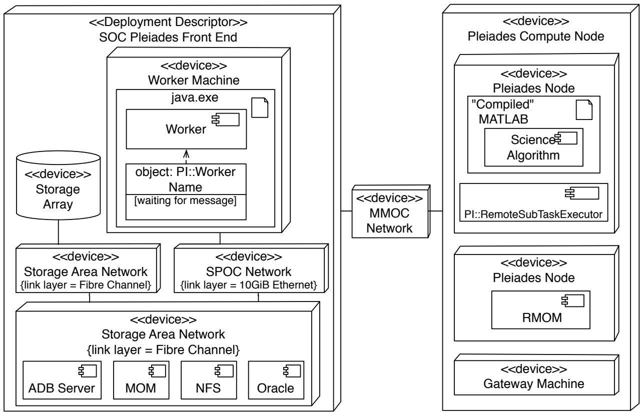

Figure 7: Pipeline deployed on Pleiades. In this deployment, the local cluster is used to generate inputs and outputs for subtasks on Pleiades. Remote processes manage the execution of science algorithm implementations. This can scale up to tens of thousands of independent subtasks.

E. Data accountability

Data accountability is an integral and crosscutting feature and is designed into the pipeline infrastructure, datastore, data receipt, and each science pipeline module. Tracked data are assigned a unique originator ID that determines its origin. PI manages the sets of parameters used for each pipeline module. These parameter sets are locked into an immutable state once a pipeline trigger has been fired. Pipeline instance IDs ensure the actual parameter values used to process any piece of data are recorded and recoverable.

F. Operating the SOC Pipeline

The data processing tasks are distributed over four operations clusters, Flight Ops, Quarterly Ops, Monthly Ops, Archive, and two test clusters, TEST and LAB. Note that all data are stored in the SAN which is partitioned to support all six clusters. All original spacecraft science and engineering data, models, algorithm parameters and configuration settings used for archival tasks are stored in the Flight Ops and Quarterly Ops partitions. Flight Ops runs PDQ and TAD to furnish pointing tweaks and target tables, which are uplinked to the spacecraft and shared with the other clusters as needed. The original data are "snapshotted" (i.e., copied on write) to Monthly Ops from Quarterly Ops to support the first pass processing on each monthly data set, whose results are used to set parameters for the final, archive processing on Quarterly Ops. The results from Quarterly Ops are snapshotted to the Archive cluster for export formatting and writing to disc, freeing up Quarterly Ops to continue processing other quarters during reprocessing activities. TEST and LAB are test clusters used for algorithm and code development, and for running science analyses, respectively.

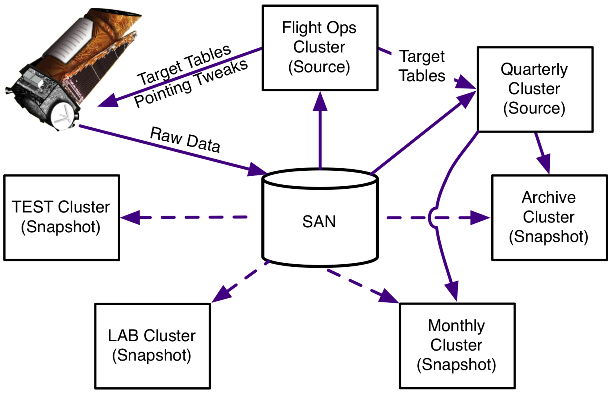
Flight Ops Cluster (Source)
Target Tables
Pointing Tweaks
Target Tables
Quarterly Cluster (Source)
Raw Data
TEST Cluster (Snapshot)
SAN
Archive Cluster (Snapshot)
LAB Cluster (Snapshot)
Monthly Cluster (Snapshot)

Figure 8: The SOC Operations Cluster Architecture

V. Conclusions

The *Kepler* Mission led to the discovery of over 2300 confirmed or validated planets, over 4700 planet candidates and over 2100 eclipsing binaries. As stunning as the exoplanet discoveries have been, the richness of the astrophysics enabled by *Kepler* may be even more breathtaking. From the detection of asteroseismic p-mode oscillations in over 15,000 stars, the detection of mixed gravity and pressure modes in red giants, the study of classical variable stars such as Gamma Doradus stars and Delta Scuti stars and hybrids, to novel astrophysics such as "heartbeat" stars and relativistic boosting [35], the science has been phenomenally diverse.

The *Kepler* SOC and science pipeline have played a principal role in generating the science data enabling these high impact science results. Developing the SOC was fraught with technical challenges both in terms of dealing with the data volume and in terms of the image and signal processing algorithms implemented in the pipeline. These challenges were met with a flexible pipeline infrastructure and transactional data base along with important algorithmic innovations, such as the multi-scale MAP approach to identifying and correcting instrumental systematic errors while retaining intrinsic astrophysical signals to the greatest degree possible. The transiting planet search components of the pipeline also saw important innovations to enable the exoplanetary science results, including the implementation of an over-complete wavelet transform-based adaptive matched filter that was coupled with χ^2 statistical vetoes to increase the discriminatory power against instrumental and non-exoplanetary transients in the flux time series data.

The success of *Kepler* and the SOC has spurred other missions such as ESA's PLATO Mission and NASA's Transiting Exoplanet Survey Satellite (TESS) Mission. In fact, the *Kepler* SOC is being retooled for use on the TESS Mission to generate the light curves and search for Earth's nearest neighbors starting in 2018 [36].

Acknowledgment

Kepler was competitively selected as the 10th NASA Discovery mission. Funding for this mission is provided by the NASA Science Mission Directorate. The author gratefully acknowledge the contributions of the greater *Kepler* team in building and operating the instrument, collecting and distributing the science data, producing the light curves and validation products employed in this publication, and archiving the results. The light curves and validation products were generated by the *Kepler* Data Processing Pipeline through the efforts of the *Kepler* Science Operations Center and Science Office. The *Kepler* Mission is led by the project office at NASA Ames Research Center. Ball Aerospace built the *Kepler* photometer and spacecraft which is operated by the Mission Operations Center at LASP. The light curves are archived at the Mikulski Archive for Space Telescopes; the Data Validation products are archived at the NASA Exoplanet Archive. The author also gratefully acknowledges the support of the NASA Advanced Supercomputing (NAS) Division within the Science Mission Directorate. The author thanks the many scientific programmers, software engineers, operations, IT support, technical writers, and management personnel who made the SOC possible.

References

[1] W. J. Borucki and et al., "Kepler planet-detection mission: Introduction and first results," *Science*, vol. 327, pp. 977–980, 2010.

[2] M. R. Haas and et al., "*Kepler* science operations," *Astroph. Journ.*, vol. 713, no. 2, p. L115, 2010.

[3] R. L. Gilliland and et al., "Kepler Mission Stellar and Instrument Noise Properties," *Astroph. Journ. Supp.*, vol. 197, p. 6, Nov. 2011.

[4] ——, "Kepler Mission Stellar and Instrument Noise Properties Revisited," *Astron. Journ.*, vol. 150, p. 133, Oct. 2015.

[5] D. A. Caldwell and et al., "Kepler instrument performance: an in-flight update," in *Space Telescopes and Instrumentation 2010: Optical, Infrared, and Millimeter Wave*, ser. Proc. SPIE, vol. 7731, Jul. 2010, p. 773117.

[6] J. M. Jenkins and et al., "Planet Detection: The Kepler Mission," in *Advances in Machine Learning and Data Mining for Astronomy*, M. J. Way, J. D. Scargle, K. M. Ali, and A. N. Srivastava, Eds. Chapman and Hall, CRC Press, Mar. 2012, pp. 355–381.

[7] D. Huber, "Precision Stellar Astrophysics in the Kepler Era," *ArXiv e-prints*, Apr. 2016.

[8] W. J. Chaplin and et al., "Ensemble Asteroseismology of Solar-Type Stars with the NASA Kepler Mission," *Science*, vol. 332, p. 213, Apr. 2011.

[9] D. Stello and et al., "Asteroseismic Classification of Stellar Populations among 13,000 Red Giants Observed by Kepler," *Astroph. Journ. Let.*, vol. 765, p. L41, Mar. 2013.

[10] S. Deheuvels and et al., “Seismic constraints on the radial dependence of the internal rotation profiles of six Kepler subgiants and young red giants,” *Astronomy and Astrophysics*, vol. 564, p. A27, Apr. 2014.

[11] K. Kolenberg and et al., “Kepler photometry of the prototypical Blazhko star RR Lyr: an old friend seen in a new light,” *MNRAS*, vol. 411, pp. 878–890, Feb. 2011.

[12] W. F. Welsh and et al., “KOI-54: The Kepler Discovery of Tidally Excited Pulsations and Brightenings in a Highly Eccentric Binary,” *Astroph. Journ. Supp.*, vol. 197, p. 4, Nov. 2011.

[13] S. Meibom and et al., “The Kepler Cluster Study: Stellar Rotation in NGC 6811,” *Astroph. Journ. Let.*, vol. 733, p. L9, May 2011.

[14] ——, “A spin-down clock for cool stars from observations of a 2.5-billion-year-old cluster,” *Nature*, vol. 517, pp. 589–591, Jan. 2015.

[15] T. C. Klaus and et al., “Kepler Science Operations Center pipeline framework,” in *Software and Cyberinfrastructure for Astronomy*, ser. Proc. SPIE, vol. 7740, Jul. 2010, p. 774017.

[16] ——, “The Kepler Science Operations Center pipeline framework extensions,” in *Software and Cyberinfrastructure for Astronomy*, ser. Proc. SPIE, vol. 7740, Jul. 2010, p. 774018.

[17] S. McCauliff, M. T. Cote, F. R. Girouard, C. Middour, T. C. Klaus, and B. Wohler, “The Kepler DB: a database management system for arrays, sparse arrays, and binary data,” in *Software and Cyberinfrastructure for Astronomy*, ser. Proc. SPIE, vol. 7740, Jul. 2010, p. 77400M.

[18] T. M. Brown, D. W. Latham, M. E. Everett, and G. A. Esquerdo, “Kepler Input Catalog: Photometric Calibration and Stellar Classification,” *Astron. Journ.*, vol. 142, p. 112, Oct. 2011.

[19] D. Huber and et al., “Revised Stellar Properties of Kepler Targets for the Quarter 1-16 Transit Detection Run,” *Astroph. Journ. Supp.*, vol. 211, p. 2, Mar. 2014.

[20] S. T. Bryson and et al., “Selecting pixels for kepler downlink,” in *Software and Cyberinfrastructure for Astronomy*, ser. Proc. SPIE, vol. 7740, 2010, p. 77401D.

[21] J. Li and et al., “Photometer performance assessment in Kepler science data processing,” in *Software and Cyberinfrastructure for Astronomy*, ser. Proc. SPIE, vol. 7740, Jul. 2010, p. 77401T.

[22] H. Chandrasekaran and et al., “Semi-weekly monitoring of the performance and attitude of Kepler using a sparse set of targets,” in *Software and Cyberinfrastructure for Astronomy*, ser. Proc. SPIE, vol. 7740, Jul. 2010, p. 77401B.

[23] J. M. Jenkins, D. J. Peters, and D. W. Murphy, “An efficient end-to-end model for the Kepler photometer,” in *Modeling and Systems Engineering for Astronomy*, ser. Proc. SPIE, S. C. Craig and M. J. Cullum, Eds., vol. 5497, Sep. 2004, pp. 202–212.

[24] S. T. Bryson and et al., “The Kepler end-to-end model: creating high-fidelity simulations to test Kepler ground processing,” in *Modeling, Systems Engineering, and Project Management for Astronomy IV*, ser. Proc. SPIE, vol. 7738, Jul. 2010, p. 773808.

[25] C. Allen, T. Klaus, and J. Jenkins, “Kepler Mission’s focal plane characterization models implementation,” in *Software and Cyberinfrastructure for Astronomy*, ser. Proc. SPIE, vol. 7740, Jul. 2010, pp. 77 401E–77 401E–8.

[26] J. M. Jenkins and J. Dunnuck, “The little photometer that could: technical challenges and science results from the Kepler Mission,” in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, ser. Proc. SPIE, vol. 8146, Sep. 2011, p. 814602.

[27] E. V. Quintana and et al., “Pixel-level calibration in the Kepler Science Operations Center pipeline,” in *Software and Cyberinfrastructure for Astronomy*, ser. Proc. SPIE, vol. 7740, Jul. 2010, p. 77401X.

[28] J. C. Smith, R. L. Morris, J. M. Jenkins, S. T. Bryson, D. A. Caldwell, and F. R. Girouard, “Finding optimal apertures in kepler data,” *Pub. Astron. Soc. Pacific*, vol. 128, no. 12, p. 124501, Dec. 2016.

[29] M. C. Stumpe and et al., “Multiscale systematic error correction via wavelet-based bandsplitting in kepler data,” *Pub. Astron. Soc. Pacific*, vol. 126, no. 935, pp. 100–114, 2014.

[30] J. C. Smith and et al., “*Kepler* presearch data conditioning II - a bayesian approach to systematic error correction of *Kepler* data,” *Pub. Astron. Soc. Pacific*, vol. 124, no. 119, pp. 1000–1014, 2012.

[31] T. N. Gautier, III and et al., “Kepler-20: A Sun-like Star with Three Sub-Neptune Exoplanets and Two Earth-size Candidates,” *Astroph. Journ.*, vol. 749, p. 15, Apr. 2012.

[32] J. M. Jenkins, “The impact of solar-like variability on the detectability of transiting terrestrial planets,” *Astroph. Journ.*, vol. 575, p. 493, 2002.

[33] J. M. Jenkins and et al., “Transiting planet search in the *Kepler* pipeline,” in *Software and Cyberinfrastructure for Astronomy*, ser. Proc. SPIE, vol. 7740, 2010, p. 77400D.

[34] J. L. Christiansen and et al., “The Derivation, Properties, and Value of Kepler’s Combined Differential Photometric Precision,” *Pub. Astron. Soc. Pacific*, vol. 124, pp. 1279–1287, Dec. 2012.

[35] M. H. van Kerkwijk, S. A. Rappaport, R. P. Breton, S. Justham, P. Podsiadlowski, and Z. Han, “Observations of Doppler Boosting in Kepler Light Curves,” *Astroph. Journ.*, vol. 715, pp. 51–58, May 2010.

[36] G. R. Ricker and et al., “Transiting Exoplanet Survey Satellite (TESS),” *J. of Astron. Telescopes, Instr., and Sys.*, vol. 1, no. 1, p. 014003, Jan. 2015.