The Chandra Source Catalog: Processing and Infrastructure

Janet D. Evans[1], I. N. Evans[1], K. J. Glotfelty[1], R. Hain[1], D. M. Hall[1], J. B. Miller[1], D. A. Plummer[1], P. Zografou[1]
C. S. Anderson[1], N. R. Bonaventura[1], J. C. Chen[1], J. E. Davis[2], S. M. Doe[1], G. Fabbiano[1], E. Galle[1], D. G. Gibbs[1], J. D. Grier[1], P. N. Harbo[1], X. He[1], J. C. Houck[2], M. Karovska[1], J. Lauer[1], M. L. McCollough[1], J. C. McDowell[1], A. W. Mitschang[1], D. L. Morgan[1], J. S. Nichols[1], M. A. Nowak[2], F. A. Primini[1], B. L. Refsdal[1], A. H. Rots[1], A. L. Siemiginowska[1], B. A. Sundheim[1], M. S. Tibbetts[1], D. W. Van Stone[1], S. L. Winkelman[1]
[1]Smithsonian Astrophysical Observatory, [2]MIT Kavli Institute

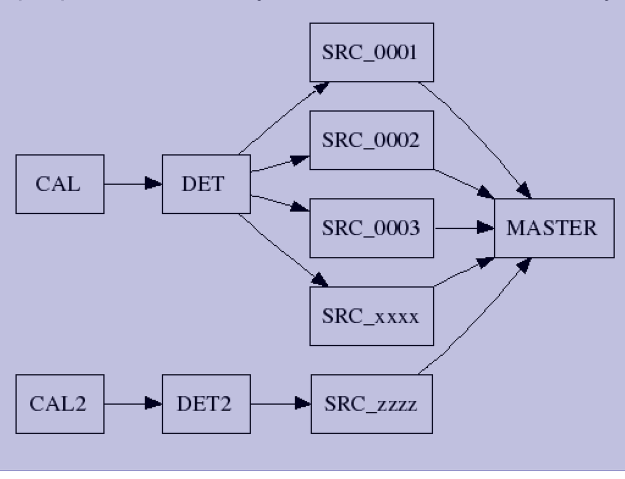
Abstract

Chandra Source Catalog processing recalibrates each observation using the latest available calibration data, and employs a wavelet-based source detection algorithm to identify all the X-ray sources in the field of view. Source properties are then extracted from each detected source that is a candidate for inclusion in the catalog. Catalog processing is completed by matching sources across multiple observations, merging common detections, and applying quality assurance checks.

The Chandra Source Catalog processing system shares a common processing infrastructure and utilizes much of the functionality that is built into the Standard Data Processing (SDP) pipeline system that provides calibrated Chandra data to end-users. Other key components of the catalog processing system have been assembled from the portable CIAO data analysis package. Minimal new software tool development has been required to support the science algorithms needed for catalog production. Since processing pipelines must be instantiated for each detected source, the number of pipelines that are run during catalog construction is a factor of order 100 times larger than for SDP. The increased computational load, and inherent parallel nature of the processing, is handled by distributing the workload across a multi-node Beowulf cluster. Modifications to the SDP automated processing application to support catalog processing, and extensions to Chandra Data Archive software to ingest and retrieve catalog products, complete the upgrades to the infrastructure to support catalog processing.

CSC Processing Pipeline

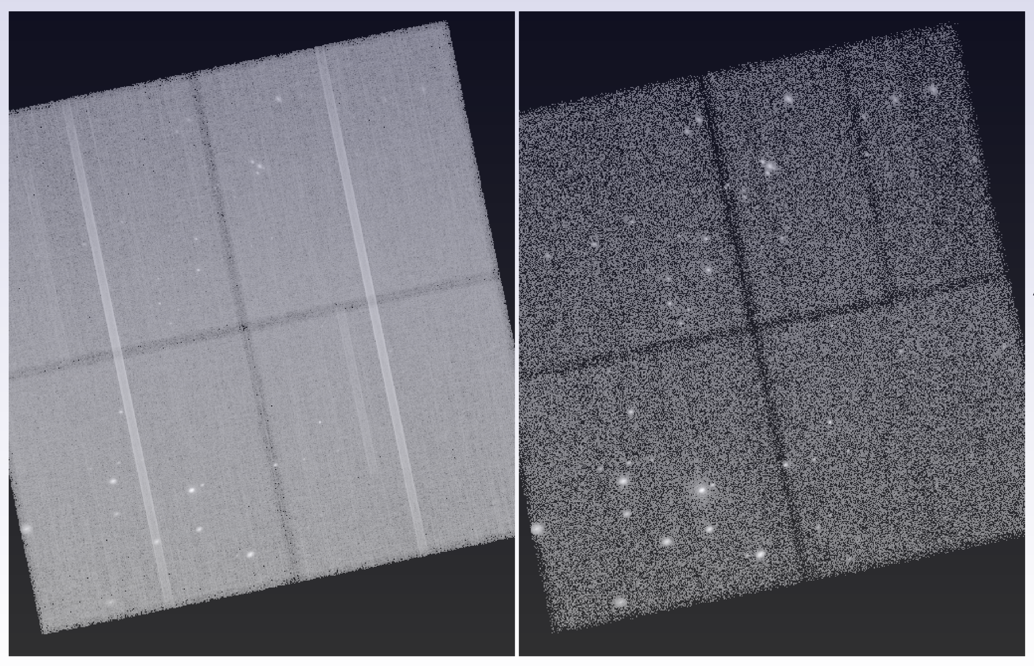
The CSC pipeline starts with re-calibration of the event data of each per-OBI observation. It is followed by a pipeline that runs source detection. For each detected source, a source pipeline is run to compute source properties. Finally sources at the same sky position are combined into master sources.



Calibrate Pipeline

For each Observation Interval:
- Remove bad pixels
- Recalibrate with "latest" Calibration data (gain, tgain, cti, degap, etc)
- Apply standard filters: Good-Time-Intervals (no aspect), ACIS grades, status bits
- Identify background flares
- Perform instrument specific steps: ACIS destreak; HRC dead time correction

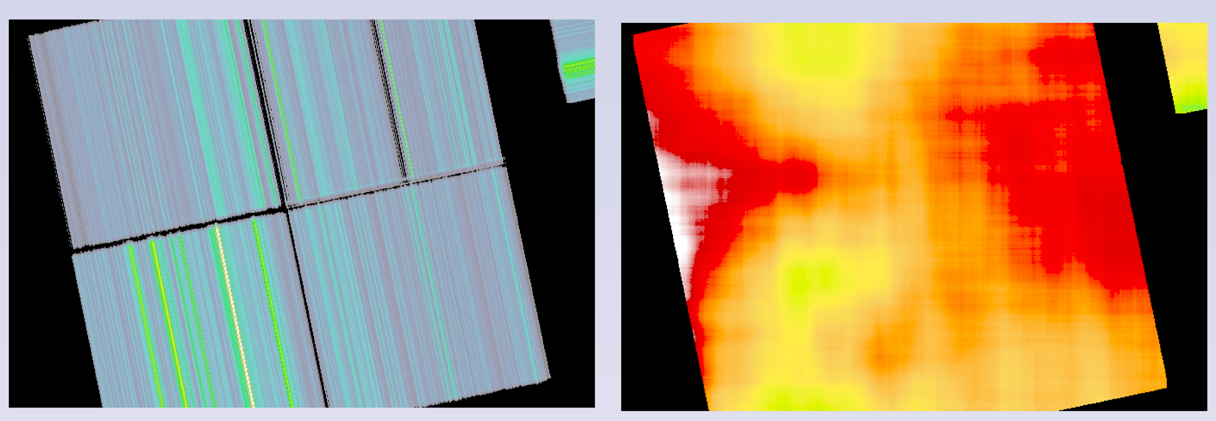
Before (L1)



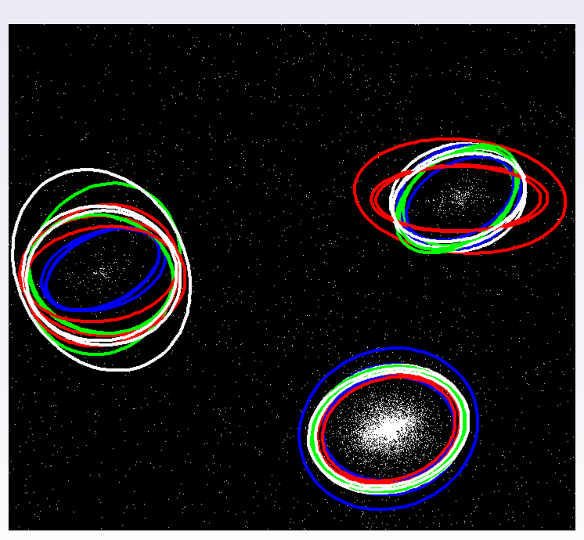
After (L3)

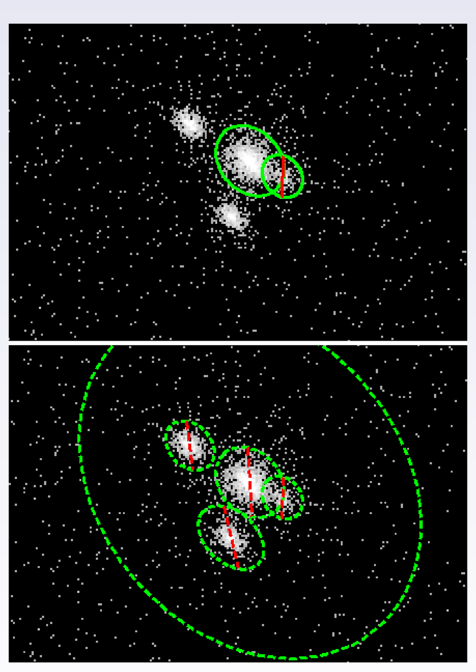
Detect Pipeline

For each Observation Interval:
- Create background map
  - HRC: use Low frequency "mean" map
  - ACIS: Split into High Frequency "Streak" map (left), Low frequency "mean" map (right)



- Run Source Detection
  - Run independently for each block/band combination
  - Uses background and exposure maps
- Combine detections (left)
- Create Source and background regions (right)
- Compute Limiting Sensitivity





Source Pipeline

Spatial Analysis
- Extract images & exposure map
- Simulate PSF w/ SAOTrace
- Measure Source and PSF size using Mexican Hat wavelet optimization
- Flag extended sources

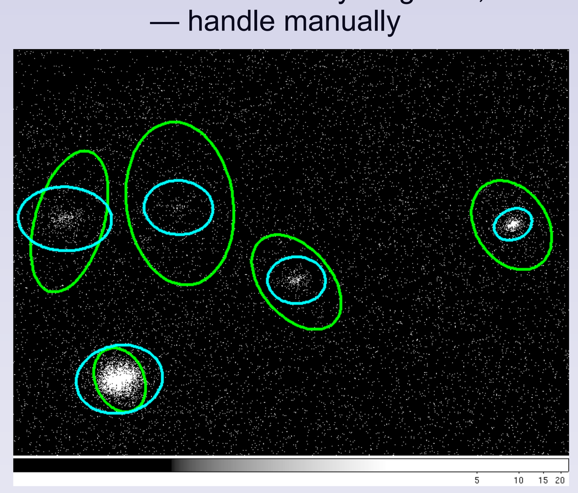
Temporal Analysis
- Reduce events to a common Good Time Interval (GTI) across all chips
- Compute time-resolved fraction of aperture area
- Compute light curve using Gregory-Loredo algorithm
- Compute Kolmogorov-Smirnov + Kuiper test probabilities of source being non-uniform

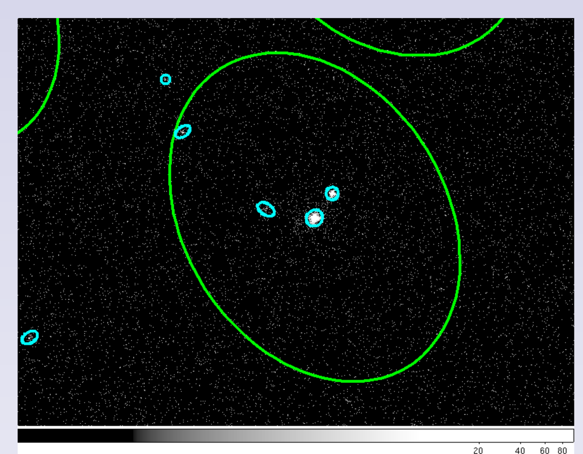
Spectral Analysis
- Compute spectrum, Ancillary Response Function (ARF), Response Matrix (RMF)
- Sources >= 150 net counts are grouped and fit with 2 spectral models
  - Absorbed black body & power law spectra

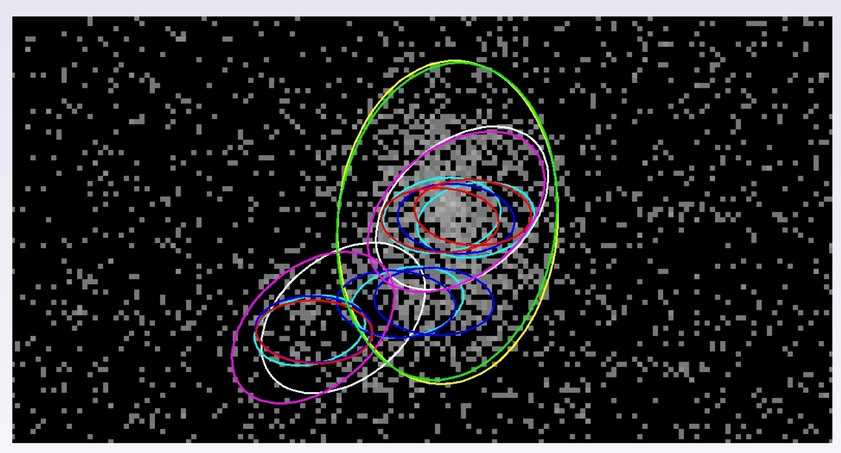
Compute Properties
- Compute position errors
- Compute 90% PSF encircled energy region
- Compute aperture photometry in source and 90% EE regions
  - per energy band using several algorithms
- Compute cross-band hardness ratios

Master Pipeline

Matching Per-Obi sources
- Spatial matching of 90% EE regions
- Determine classification:
  - New — doesn't overlap sources in any other OBIs
  - Simple — all overlapping sources are unique within an OBI
  - Confused — 1 source in one OBI overlaps more than one source in another obi
  - Too hard — everything else, 'chained sources', etc — handle manually



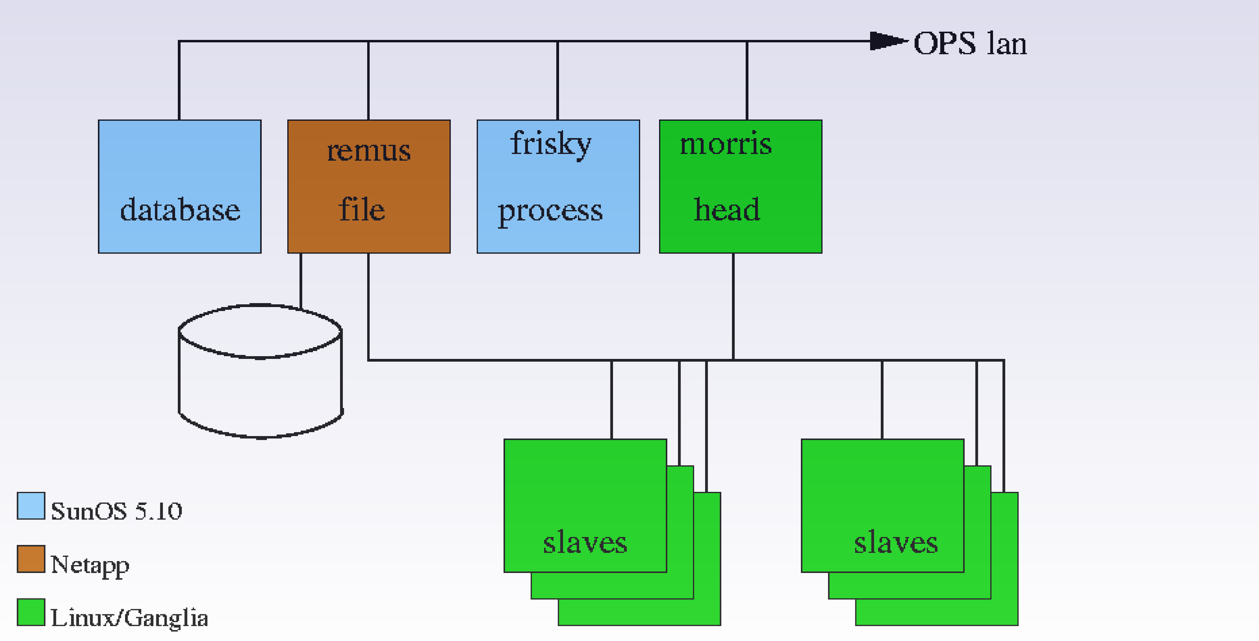




- Merging properties
  - Simple rules in Release I
  - Compute true inter-obi values
    - variability
    - positions & errors
    - aperture photometry
  - Pick "the best" OBI values for most other properties
  - Package DB instructions

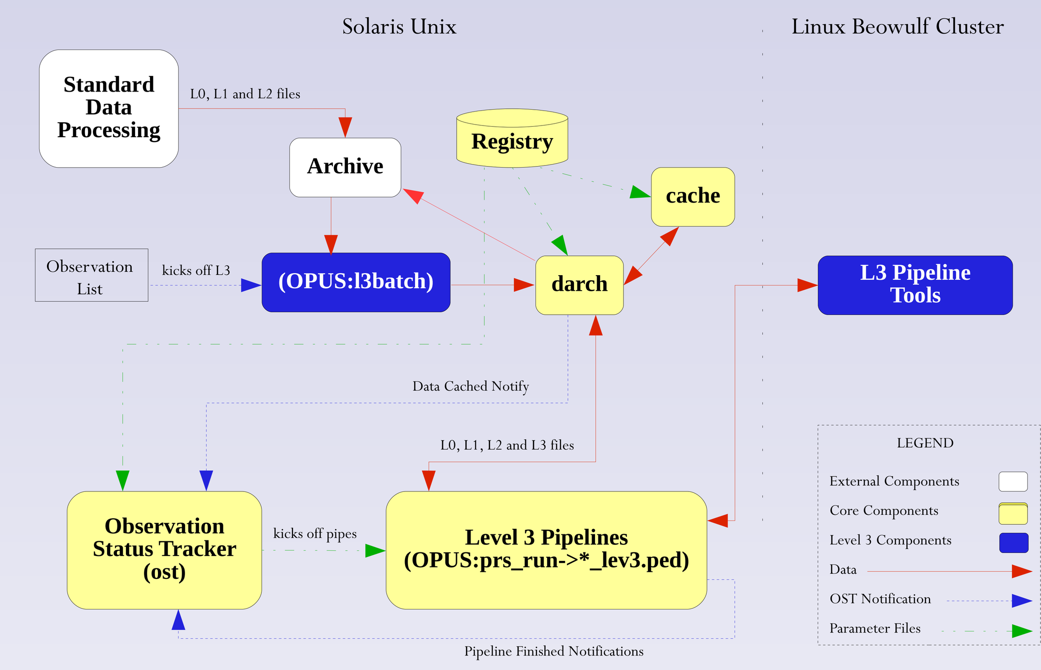
Hardware Architecture

The fundamental difference between Level 3 processing and SDP is that science pipelines are instantiated for each detected source, rather than for each observation. Consequently, the number of pipelines that are run in Level 3 processing is a factor of order 100 times larger than in SDP. The increased computational load, and inherent parallel nature of the processing, is being handled by modifying the processing infrastructure to distribute the workload across the nodes of a multi-node Beowulf cluster.



Automated Processing Application (AP)

AP is a general template-driven automated data processing system that was designed to adapt to different data processing requirements. Standard Data Processing (SDP) is the AP system that processes current Chandra data on a daily basis. L3AP utilizes the core AP/SDP components (with a few updates) along with a layer that handles L3 specific requirements. AP utilizes OPUS, a generic distributed pipeline system developed by the Space Telescope Science Institute. See http://www.stsci.edu/resources/software_hardware/opus.



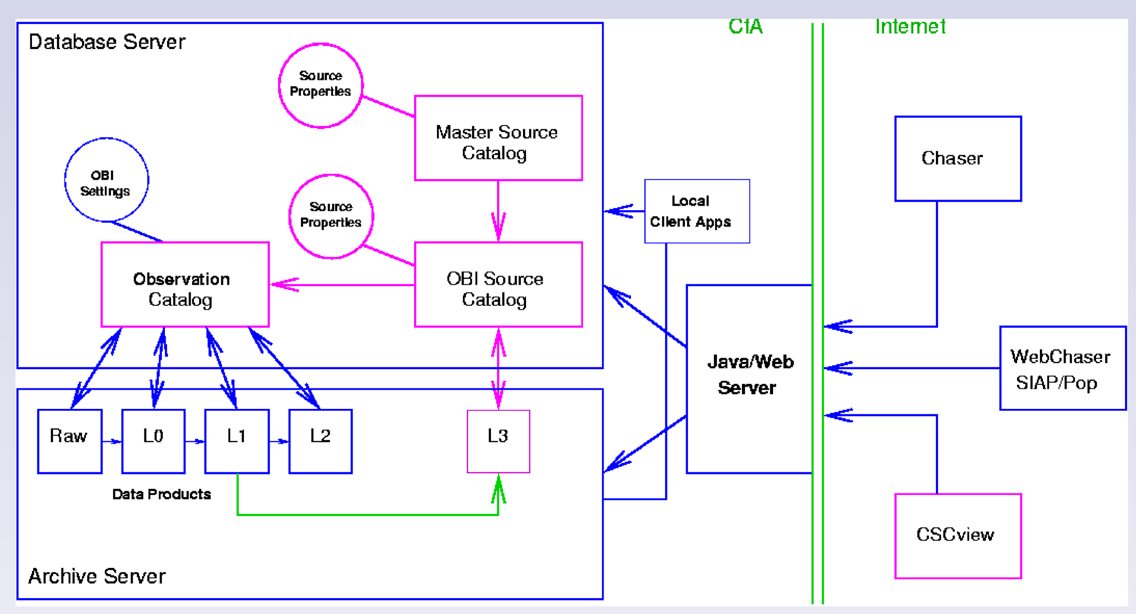
Core AP Components:
- cache - data store containing pipeline input/output data
- darch - data archiver/retriever - provides an interface to the archive (utilizes the data cache)
- Observation Status Tracker (OST) - tracks data and initiates pipelines when appropriate
- OPUS: prs_run - OPUS task that is executed for each Chandra pipeline. It executes applications in stages to retrieve the required data from cache via darch, executes the pipeline tools and stores the results to cache for use by other Chandra pipelines.
- Registry - provides a flexible way to define data products, processing pipelines, pipeline kickoff criteria and data dependencies between pipelines. It orchestrates automated processing via ASCII parameter files that are read by the core AP components.

Level 3 Components:
- OPUS: l3batch - An OPUS task that seeds the data cache with L3 input data and monitors the L3 pipelines for each input observation.
- L3 Pipeline Tools - applications that are run as part of a pipeline on the Beowulf Cluster and produce the L3 data products

Core AP Component updates necessary for L3 processing:
- The processing speed necessary for L3 resulted in some updates to the core AP components.
- The Registry and prs_run task needed to be updated to accommodate running pipelines on the Linux Beowulf cluster.
- OST and cache_server needed to be updated to be more efficient to accommodate the high data flow rates.

Chandra Data Archive (CDA)

The CSC is incorporated in the CDA with the addition of databases and L3 products storage on dedicated CSC servers within the existing archive architecture. Lower level databases that are tightly coupled with the CSC are replicated to the CSC servers for easier access.



The *arc4gl* archive server language is extended with L3 options for access of the data products. A new web-based user interface (CSCview), has been developed using the existing Java/Web middle tier, for browsing of the databases and data retrieval.

The CSC databases store all versions of sources and implement many-to-many relationships between master sources and sources-by-observation. Source-by-observation records are also linked to their observation records in OCAT. In order to meet the multi-version source requirement, the OCAT is enhanced to store multiple versions of data fields linked to the CSC. All versions of L3 data products are kept in the archive and have links to their associated source-by-observation versions.

In order to support periodic releases and snapshots, the CSC database records have *creation time stamp* and *catalog release label*. A user view of the catalog selects by label to see a release or by time for a snapshot. In both cases, the view selects qualified master sources, then follows the versioning links to sources-by-observation and associated data products.

ACKNOWLEDGEMENTS

Support for development of the CSC is provided by the National Aeronautics and Space Administration through the *Chandra X-ray Center*, which is operated by the Smithsonian Astrophysical Observatory for and on behalf of the National Aeronautics and Space Administration under contract NAS 8-03060.