Progress Report

Ian Evans (CXCDS) and Jonathan McDowell (SDS)

Chandra Users' Committee Meeting

September 19, 2007

Outline

Acknowledgements

- The *Chandra Source Catalog* is a collaborative effort between the scientists and software developers of CXCDS and the scientists of SDS
- This presentation includes contributions from both groups
  - Other CXC scientists have provided contributions in some areas

Presentation Outline

- This presentation is divided into two parts
  - Outline of overall progress since the last CUC meeting and schedule update

    Presented by Ian on behalf of the *Chandra Source Catalog* team (CXCDS and SDS)
  - Detailed discussion of science progress, issues, and investigations, with examples

    Presented by Jonathan on behalf of the *Chandra Source Catalog* team (SDS and CXCDS)

Summary

Summary

- Significant progress has been made since the last CUC meeting
  - About a dozen requirements and specs have been delivered by the science team
  - A similar number are currently in-work, including several that have science studies in progress or completed
  - The software team has completed the CAT 2.5 (Operational Testbed 1) release, which is now operational and providing feedback to the science and software teams
  - Development work for the CAT 2.6 (Operational Testbed 2) release is well underway
- The science and software teams are fully engaged and understand the priorities and schedule
  - Priorities are clearly established by the CXC management

Currently estimating start of production in May 2008, with first public data access in June, and formal release 1 in October

  - First public access includes ~1/3 of public imaging observations processed and available, and preliminary statistical characterization of catalog properties
  - Catalog release includes public mission-to-date imaging observations and complete statistical characterization of catalog properties available
  - This schedule reflects an approximately 2 month slip since March
    - Due to a mix of internal (catalog) and external (non-catalog) factors
    - We used the risk mitigation process defined earlier this year to limit the impacts of the slip and ensure that the teams are focused on the project goals

Progress: Science Highlights

Science Highlights Since Last CUC Meeting

- *Details will be discussed by Jonathan later in this presentation*
- Algorithms, requirements, and specifications (with targeted software releases)
  - Aperture photometry
    - Developed algorithms to compute aperture photometry rates, fluxes, and errors, and limiting sensitivity maps (CAT 2.5, 2.6)
  - Background Maps and Source Detection
    - Refined algorithms to compute ACIS streak map and low freq. background (CAT 2.5)
    - Developed algorithm to compute HRC-I low frequency background (CAT 2.6)
    - Developed specification for estimating source position uncertainties (CAT 2.5)
  - Spatial Properties
    - Developed an algorithm for identifying observations with large extended sources [to be excluded from processing] (CAT 3.0)
  - Spectral Properties
    - Developed specifications for computing spectral fits & energy flux in bands (CAT 2.5, 2.6)
    - Developed specification for computing hardness ratios (CAT 2.6)
  - Temporal Properties
    - Developed specifications for evaluating time variability of sources (CAT 2.5, 2.6)
    - Developed an algorithm for identifying and removing background flares (CAT 2.7)
  - Master Source Properties
    - Developed algorithm for combining error ellipses from multiple observations that include the same source (CAT 2.6)

Progress: Science Highlights (cont.)

Science Highlights Since Last CUC Meeting (continued)

- Science Studies
  - Background Maps and Source Detection
    - Evaluated impacts of HRC-S detector structure on background map computations
    - Evaluated performance of **`wavdetect`** on extended sources and systematic errors in source region parameters computed by **`wavdetect`**
  - Spatial Properties
    - Evaluating a scale-less algorithm for estimating source extent
  - Temporal Properties
    - Evaluating approaches for removing effects of dither from time variability measures
  - Master Source Properties
    - Developed estimator of pile-up fraction as a function of count rate

Documents

Documents Delivered Since Last CUC Meeting

- Requirements and specifications
  - "Combining Error Ellipses", J. Davis, SDS/MIT
  - "*Chandra Source Catalog* Quality Assurance Specifications", I. Evans, CXCDS
  - "*Chandra Source Catalog* Requirements, Version 0.6", I. Evans, CXCDS
  - "Computation of Hardness Ratios Using BEHR", I. Evans, CXCDS
  - "*Chandra Source Catalog* Background Map Flux, Edge, and Exposure Map Modifications", M. McCollough & A. Rots, CXCDS
  - "HRC-I Background Map Requirements", M. McCollough & A. Rots, CXCDS
  - "Avoiding Numerical Instabilities in Computing Aperture Fluxes", F. Primini, SDS
  - "Computing L3 Limiting Sensitivity Maps", F. Primini, SDS
  - "Recommendations for Estimating L3 Source Position Uncertainties", F. Primini, SDS
  - "Revised Specifications for Computing Aperture Photometry Quantities", F. Primini, SDS
  - "Detection of Background Flares and Similar Phenomena", A. Rots, CXCDS
  - "Identification of Observations Containing Large Extended Sources", A. Rots, CXCDS
  - "L3 Temporal Variability Measures", A. Rots, CXCDS
- Science study reports
  - "Pile-up Fractions and Count Rates", J. Davis, SDS/MIT
  - "Measuring Detected Source Extent Using Mexican-Hat Optimization", J. Houck, SDS/MIT
  - "HRC-S Background Map Issues", M. McCollough & A. Rots, CXCDS
  - "Preliminary Assessment of Dither Corrections for L3 Lightcurves", M. Nowak, SDS/MIT
  - "Thoughts on Filtering Flares in L3 Background Files", M. Nowak, SDS/MIT
  - "Evaluation on wavdetect Performance on Extended Sources", F. Primini, SDS

Software Highlights Since Last CUC Meeting

- Requirements and specifications
  - Revised *Chandra Source Catalog* Requirements document to version 0.6
    - Major change is the addition of two-sided confidence intervals for photometric and spectrometric quantities
      - Aperture fluxes and associated values, spectral fit parameters, hardness ratios
      - Replaces symmetric Gaussian errors in previous version
    - Updates to definition of temporal variability properties
      - Now record several quantities that are already calculated in algorithm
    - Improved definitions of the contents of some data products
      - Removed TBDs, added clarifications based on science use cases and requirements for calculating limiting sensitivity
    - A number of minor changes/clarifications, and resolution of several TBDs
  - Quality assurance requirements updated
    - Incorporates more robust false source rejection tests
    - Details requirements for manual quality assurance steps and GUI
    - Feedback from CAT 2.5 testing valuable for assessing quality assurance efficacy
  - Catalog infrastructure reviews
    - Extensive reviews of quality assurance, catalog inclusion criteria, master pipeline transactions, and database interfaces established detailed system-wide infrastructure

Progress: Software Highlights (cont.)

Software Highlights Since Last CUC Meeting (continued)

- CAT 2.5 (Operational Testbed 1) build released
  - First release with archive and database integration with processing pipelines
  - Pipelines match Requirements Document version 0.5, with waivers for incomplete items, and some updated data products (per draft version 0.6)
    - *Calibrate*, *Detect*, and *Source* pipelines
      - Source detection with combined streak map and low freq. backgrounds for ACIS
      - Aperture photometry rates, fluxes, and errors
      - Sensitivity map and spectral fits (initial implementations)
      - Gregory-Loredo, KS, and Kuiper variability and light-curve (source region only)
      - PSF generation running directly on Beowulf using SAOTrace
      - No source extent estimation
    - Initial version of *Master* pipeline
      - Matches sources across multiple observations to create a single "top-level" catalog entry for each source
      - Creates database transaction files
      - Does not include merge properties computations
    - No quality assurance
  - Automated Processing (AP) infrastructure to run system in "catch-up" mode
  - Archive
    - Observation and master source databases
    - Database quantity and archive data product ingestion
    - AP/archive interfaces

Results from CAT 2.5 Evaluation

Progress

- CAT 2.5 running and providing feedback to science and software teams
  - Approximately 50 observations processed to develop a baseline
    - Chosen to exercise various aspects of the algorithms, for comparison with existing studies, and to include a range of sources, source distributions, and densities
  - Additional observations will be run for performance evaluation and to expand the source sample for statistical analysis
- *Science results will be reported by Jonathan later in this presentation*

Issues

- Science algorithms
  - Main issue to be addressed is the high false source rate for some observations
    - Limitations of ACIS readout streak computation for subarrays with few (< 256 ?) rows
      - Reduced statistics results in improper estimate of streak background
      - Investigating use of background variance in **`wavdetect`** step, as well as quality assurance flagging to eliminate false sources
    - HRC low frequency background improperly determined
      - Background estimate does not exclude the "bad" outer regions of the detector
      - Expect to resolve on production timescale (not necessarily start of production)
    - HRC-S LESF results in false source detections
      - Need additional mechanism to consider spatial scale for background variability at edge of LESF
      - *Do not expect to resolve for release 1; current baseline is to exclude HRC-S*

Results from CAT 2.5 Evaluation (cont.)

Issues (continued)

- Software
  - Initial integration of pipelines with archive and database took longer than expected
    - Several bugs and performance issues were addressed
  - Needed several iterations to get source detection working correctly with computed backgrounds
    - Pipeline parameters updated based on analysis of a sample of observations
    - Lack of quality assurance steps required workarounds to eliminate propagation of false source detections (source regions with low net counts and significance)
- Hardware performance and reliability
  - Archive test hardware performance issues
    - Will be addressed as part of archive hardware upgrades currently in work (to be installed prior to start of production)
  - Beowulf cluster hardware reliability
    - The first time all cluster processors were run at a sustained ~100% utilization during CAT 2.5 testing, one node (out of 14) suffered an unrecoverable electrical failure
    - Other nodes occasionally crash unexpectedly, leaving pipelines hanging
    - Performance about a factor ~2 slower than expected
    - Plan to replace existing cluster nodes with new quad-core nodes prior to start of production

Schedule: Planned Software Releases

Planned Software Releases

- CAT 2.6 (Operational Testbed 2; November 2007)
  - Pipelines Match Requirements Document version 0.6, with waivers for incomplete items
  - *Calibrate*, *Detect*, and *Source* pipelines essentially complete
    - Exceptions are some source extent properties and lack of background flare filtering
    - Includes automated pipeline quality assurance
      - Validate processing, detected sources (false source rejection), source properties
  - *Master* pipeline mostly complete
    - Most merged source properties
    - Inter-observation variability
    - No quality assurance
  - Automated Processing infrastructure
    - Support for automated quality assurance
    - Support for "bulk" reprocessing (observation at a time)
  - Archive
    - Versioning evaluation with reprocessing integration
    - Includes quality assurance/catalog inclusion transactions

Schedule: Planned Software Releases (cont.)

Planned Software Releases (continued)

- CAT 2.7 (Production Prototype; January 2008)
  - All pipelines in production configuration
    - Completion of any remaining production liens
  - Automated Processing infrastructure
    - Mechanism for selective reprocessing (*e.g.*, failed pipelines)
    - Support manual quality assurance
  - Manual quality assurance GUI
    - Evaluate failed pipelines
    - Evaluate and correct questionable source detections and source matches
  - Archive
    - Initial user interface GUI
    - Catalog release support (freeze catalog and evaluate catalog inclusion criteria)
  - Run pre-production test
- CAT 3.0 (Production Release; May 2008)
  - Pipelines
    - Fine tune pipelines based on pre-production test
  - Automated Processing infrastructure
    - Operations interfaces complete; support for daily operations
  - Archive
    - User interface GUI
    - Hardware configurations complete
  - Catalog production

Schedule: Tasks Leading to Catalog Production

Tracked Task Summary

- High level science and software tasks leading to production start
  - Does not include science characterization tasks which occur in parallel, since they are not a lien against production start

	Release	Completed	Working	Waiting*	Open	Hold*
CAT 2.5 Science		21				
CAT 2.5 Software	Aug 07	36				
CAT 2.6 Science		7	3	4		1
CAT 2.6 Software	Nov 07	4	19		17	1
CAT 2.7 Science		6	5		7	
CAT 2.7 Software	Jan 08	1	2		22	
CAT 3.0 Science		6	3		5	
CAT 3.0 Software	May 08	1	2		12	

* "Waiting" tasks are in-work but require feedback from CAT 2.5 evaluation to complete
"Hold" tasks are suspended (identified as not required for production start)

Schedule: Risks

Schedule Risk Changes Since Last CUC Meeting

- Processing and/or archive hardware performance is inadequate (Risk ↓↓)
  - CAT 2.5 testing indicates upgrades required to for aging Beowulf hardware to achieve performance goals
    - Current pipelines include all "compute-intensive" steps, so no further significant performance changes are anticipated
    - Hardware plan being formulated to recover performance, and will be ordered shortly
  - Database and archive performance evaluation has established requirements for archive hardware upgrades
    - These upgrades have been folded into the overall archive hardware migration plan, and new hardware has been ordered
- Science algorithm development takes longer than planned (Risk ↓)
  - Development of key algorithms is now completed or well in hand
    - Most are already implemented in CAT 2.5, and will be fine-tuned in future releases
  - Certain science capabilities will be deferred to avoid impacting release 1 schedule
- Science algorithms or software are inadequate (Risk ↓)
  - Each software release is followed by a test and evaluation period to identify issues early
    - CAT 2.5 testing indicates that there are some issues to be resolved but no major problems
    - Science characterization of results builds confidence
  - Pre-production test will evaluate scientific correctness on ~1/3 of the mission data prior to production start

Schedule: Risks (cont.)

Schedule Risks (continued)

- Software implementation takes longer than planned (Risk ↔)
  - Specifications or partial specifications exists for most tasks, and schedule estimates are fairly reliable for these
  - Remaining tasks have incorporated best estimates based on available information
- Establishing catalog operations takes longer than planned (Risk ↔)
  - Plan to provide system documentation and training for the operations group well in advance of pre-production test, and use pre-production test to refine procedures

    Software and science teams will support operations during transition phase
- Other tasks compete for resources (Risk ↑)
  - Same science and software resources support live *Chandra* mission
  - Some unexpected CXC software development requirements for Cycle 10 have been identified, but we expect these to have limited schedule impact
  - New archive hardware is not supported under Solaris 8, forcing us to migrate the data system to Solaris 10 by the end of the year
  - Sybase has announced it will discontinue support for the version of the SQL database server that we are running in December 2007, forcing us to upgrade to a new version

    Our plan is to combine the migration to Solaris 10 and the Sybase 15 upgrade in order to minimize the resulting schedule impact

Schedule: Summary

Schedule

- Continuing to maintain a detailed task list and schedule leading to catalog release 1
  - All of the tasks (science and software) that must be done for each software release are expected to be completed in time to avoid delaying that release
    This is true even for some tasks that are presently lagging
  - Some tasks have been completed ahead of schedule
- Several of the schedule risks have decreased since the last CUC meeting, as more of the higher-risk, critical tasks are completed
- Most concerned about schedule risk associated with competition for available resources
  - Some unanticipated external factors do require action by the end of this year
  - Will work to mitigate by reassigning resources and internally shuffling schedule
- Continue to estimate that all catalog schedule components will complete at about the same time

THERE ARE STILL NO LONG POLES

Bottom line best estimate is start of production in May 2008, with first public data access in June, and formal release 1 in October