Astronomical Data Analysis Software and Systems VII
ASP Conference Series, Vol. 145, 1998
R. Albrecht, R. N. Hook and H. A. Bushouse, eds.

OPUS-97: A Generalized Operational Pipeline System

J. Rose

CSC: Computer Sciences Corporation, Inc.

Abstract. OPUS is the platform on which the telemetry pipeline at the Hubble Space Telescope Science Institute is running currently. OPUS was developed both to repair the mistakes of the past, and to build a system which could meet the challenges of the future. The production pipeline inherited at the Space Telescope Science Institute was designed a decade earlier, and made assumptions about the environment which were unsustainable.

While OPUS was developed in an environment that required a great deal of attention to throughput, speed, efficiency, flexibility, robustness and extensibility, it is not just a "big science" machine. The OPUS platform, our baseline product, is a small compact system designed to solve a specific problem in a robust way.

The OPUS platform handles communication with the OPUS blackboard; individual processes within this pipeline need have no knowledge of OPUS, of the blackboard, or of the pipeline itself. The OPUS API is an intermediate pipeline product. In addition to the pipeline platform and the GUI managers, the OPUS object libraries can give your mission finer control over pipeline processing.

The OPUS platform, including a sample pipeline, is now available on CD-ROM. That package, designed to be run on the Solaris operating system, can help you decide whether OPUS can be used for your own mission.

OPUS was developed in an environment which demanded attention to productivity. A ten-fold increase in the volume of data was anticipated with the new instrumentation to be flown on the Hubble, and the ground system required a pipeline which could accommodate that load. OPUS has met these challenges.

A distributed pipeline system which allows multiple instances of multiple processes to run on multiple nodes over multiple paths may seem like an operational nightmare. To resolve this, OPUS includes two pipeline managers: Motif GUI applications which assist operations staff in monitoring the system. The Process Manager not only assists with the task of configuring the system, but monitors what processes are running on which nodes, and what they are doing currently. The Observation Manager provides a different view of the pipeline activities, monitoring which datasets are in which step in the pipeline and alerting the operator when observations are unable to complete the pipeline.

[1]rose@stsci.edu

The success of OPUS can be attributed in part to adopting a blackboard architecture of interprocess communication. This technique effectively decouples the communication process and automatically makes the entire system more robust. Based upon the standard file system, OPUS inherits a simple, robust and well-tested blackboard.

OPUS has been operational at the Space Telescope Science Institute since December 1995, and has now been packaged[2] so that other missions can take advantage of this robust, extensible pipeline system.

1. Fully distributed processing...

OPUS supports a fully distributed processing system. This means multiple instances of a process are able to be run simultaneously without interference from one another. In addition it can support a variety of processes, each a step in the pipeline.

Moreover multiple pipelines, or paths, are supported. For example, at the Space Telescope Science Institute it is necessary to operate a real-time pipeline at the same time that a production pipeline is processing. And a reprocessing pipeline may be simultaneously converting science images in the background.

In addition to several pipelines with identical processing steps, OPUS supports any number of distinct pipelines all running on the same set of processors. Thus, in addition to the science pipelines, OPUS accommodates an engineering data pipeline, a separate pipeline for other non-science data, as well as an interface to Goddard Space Flight Center for data receipt.

All pipelines are defined in simple text files. The pipeline path file defines a set of network-visible directories on the shared disks. While one set of disks is being used to process one kind of data, another set can be employed to process a different type of data. ASCII text files are the basis for configuring any component of the OPUS environment. Adding an additional machine to the set of nodes is accomplished by editing a text file: that node will immediately be available to share the load.

The OPUS managers are uniquely suited to keep track of the environment. The Process Manager keeps track of what is running where, while the Observation Manager is monitoring the progress of observations being processed in any one of the pipelines. Multiple Observation Managers can each monitor their own pipelines without interference from one another.

2. ...or a simple data pipeline

Even though OPUS was developed in an environment that required a great deal of attention to throughput, speed, efficiency, flexibility, robustness and extensibility, it is not just a "big science" machine. The OPUS platform, our

[2]http://www.stsci.edu/opus/

baseline product, is a compact[3] system designed to solve a specific problem in a robust way.

OPUS is implemented as an automated pipeline, one which can start up automatically, send exposures from task to task automatically, and monitor how things are proceeding automatically. OPUS is designed to achieve a "lights-out" operation: data can enter the pipeline, be processed, then archived without intervention.

The OPUS pipeline managers monitor the status of each exposure in the system: how far it got in the pipeline, whether it failed and where it failed. The GUI interfaces provide convenient tools to investigate problems, examine log files, view trailers, and restart troubled exposures at any step in the pipeline.

The FUSE (Far Ultraviolet Spectrographic Explorer) team selected OPUS for their pipeline even though they will be receiving only a moderate amount of data that will processed on a single workstation. OPUS was chosen because it frees the FUSE team to concentrate on the science and calibration issues which are unique to their instrument.

By handling the mechanics of data processing, OPUS frees the scientists to do science.

3. OPUS-97 platform

The OPUS platform is the baseline pipeline product. All communication with the OPUS blackboard is handled by the platform; an individual process within this pipeline need have no knowledge of OPUS, of the blackboard, or of the pipeline itself.

OPUS can accommodate any non-interactive shell script. When there is work to be performed by that script, OPUS will pass the name of the dataset to be processed, the location of that dataset and other auxiliary datasets, as well as other parameters required by the script.

Similarly the OPUS platform can wrap any stand-alone, non-interactive executable which takes the name of the input dataset as a single argument. All other information for that task is either passed by OPUS through environment variables (symbols) or is obtained from the dataset itself.

OPUS is fully table-driven. To add another process to the OPUS pipeline only requires the development of an ASCII text file describing the command line arguments, the pipeline triggers, subsequent process triggers, and other such control information.

Processes (or scripts) can be triggered in three ways: the most common way is to allow the completion of one or more previous pipeline steps to act as the process trigger mechanism. Another useful technique is to use the existence of a file as the trigger. Alternatively one can use a time event to trigger an OPUS process (eg: wake up once an hour).

---

[3]The OPUS baseline system is less than 20,000 lines of code.

The OPUS platform is being distributed now on CD-ROM [4] for the Solaris platform. This distribution comes with a sample pipeline which shows how to set up the system and how to modify it to reflect your own needs.

4. An OPUS Pipeline

Where the OPUS platform is a generalized pipeline environment, it does not provide the applications which handle mission-specific data. The OPUS team at the STScI has a variety of tools and packages at hand to help build functional telemetry pipelines. Packages to handle FITS files, keyword dictionaries, database access, keyword population, message handling, and the like, form the building blocks of a standard astronomical data processing pipeline.

Certainly the specific applications are not portable, but the experience of the OPUS team in developing complete pipelines for Hubble and for FUSE can be easily applied to other missions.

Is OPUS overkill for a small mission? No. First, OPUS is not a large system. It is compact, designed to solve a specific problem in a robust way: distributed processing with controlled monitoring. Second, OPUS does the controlled monitoring. Telemetry processing does not have to be a labor intensive task. OPUS relieves your talented engineering and science staff to do more interesting work. Third, OPUS exists. Your mission is to understand the science, not to build pipelines.

Acknowledgments. The entire OPUS team at the Space Telescope Science Institute was involved in the development of OPUS-97: Daryl Swade (CSC), Mary Alice Rose (AURA), Chris Heller, Warren Miller, Mike Swam and Steve Slowinski are all to be congratulated.

References

Rose, J., Choo, T.H., Rose, M.A. (1995) "The OPUS Pipeline Managers" in *ADASS V*, pp311-314.

Rose, J. et al (1994) "The OPUS Pipeline: A Partially Object-Oriented Pipeline System" in *ADASS IV*, pp429-432

Nii, H.P. (1989). "Introduction" in *Blackboard Architectures and Applications*, Jagannathan, V., Dodhiawala, R., Baum, L., editors, Academic Press, San Diego, CA, pp xix-xxix.

[4] http://www.stsci.edu/opus/opusfaq.html