A wide-field images processing suite

TERAPIX (Traitement Élémentaire, Réduction et Analyse des PIXels de megacam) is an astronomical data reduction centre dedicated to the processing of extremely large data flows from digital sky surveys. Located at the IAP (Institut d'Astrophysique de Paris), its primary tasks are :

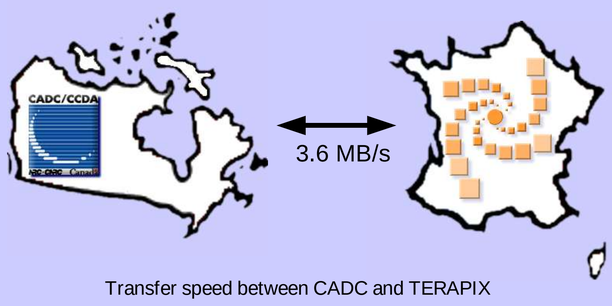
- to develop image processing and pipeline software for MegaCam (the new giant CCD camera of the CFHT telescope in Hawaii);
- to develop and provide tools for handling of large CCD images ;
- to operate the final reduction pipeline to produce calibrated images and catalogues of MegaCam images over the next 5 years ;
- to provide technical assistance and TERAPIX computing facilities to MegaCam users.

This poster presents the different components of the processing suite, chained as represented besides.

Snooppix → DataTransfer → Spica → Snooppix

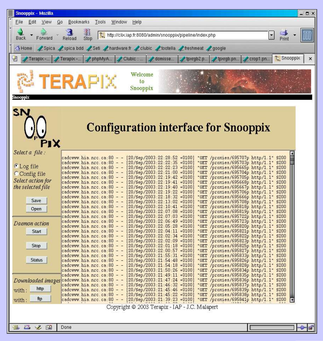
Snooppix

Perl daemon that downloads automatically data via a http or ftp protocol.

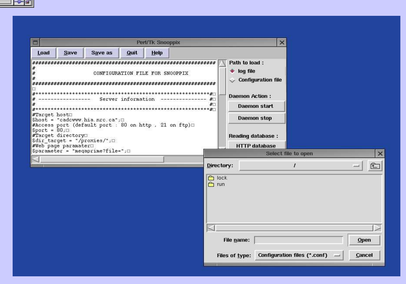


Transfer speed between CADC and TERAPIX

Based on wget (free tool for non-interactive files download from the web), Snooppix can scan a web page and download data sending simultaneous wget on each web page's link.

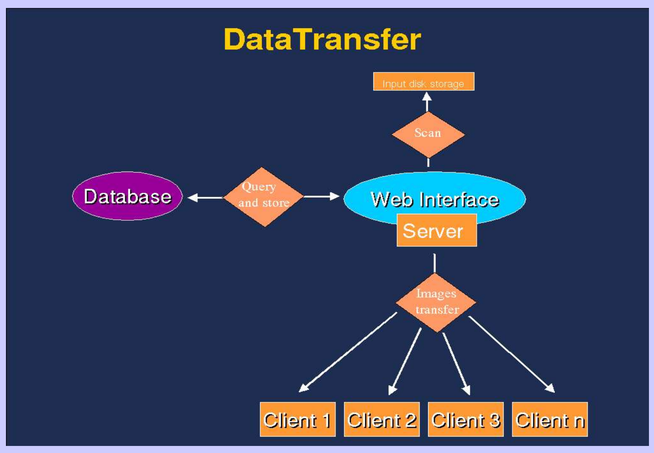


In order to avoid duplication, Snooppix stores in a DBM file the downloaded filenames. A log file is also available to check download status. Snooppix can be managed from web or perl/TK interface.Snooppix uses a configuration file to load user preferences. This file describes snooppix's connection parameters about server (web or ftp) and informations about client (directory where are downloaded data, ...).

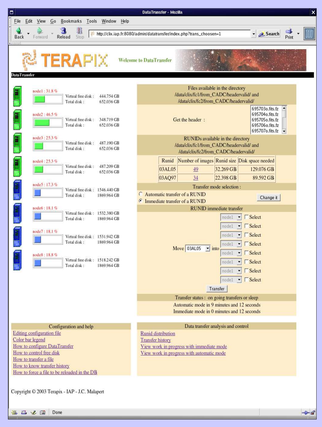


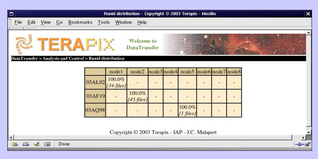
DataTransfer

Set of perl and php scripts that dispatches data and managing transfers on a cluster.



In order to transfer and sort out quickly images coming from CADC on our cluster, we have developped a software based on php and perl scripts which allows to dispatch data across a cluster.



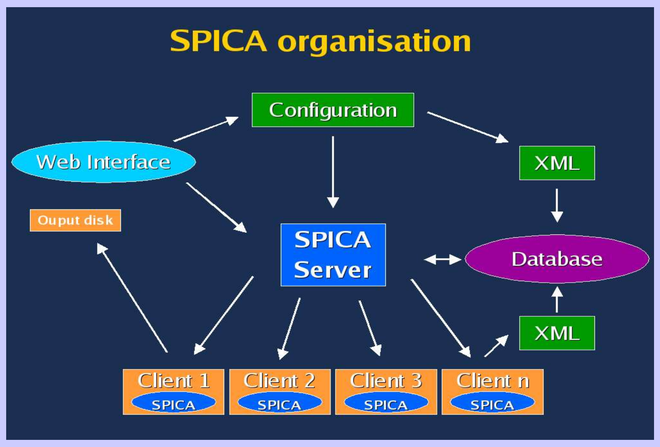


DataTransfer is managed from a web page. So, you can transfer images by simple clicks. DataTransfer is not only a tool for data transfer, it can also display :

- FITS header
- files distribution across nodes.
- transfer history
- free disk space on each nodes

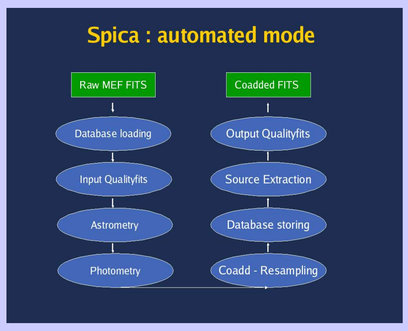
Spica

Processing tools to produce calibrated data from large astronomical images



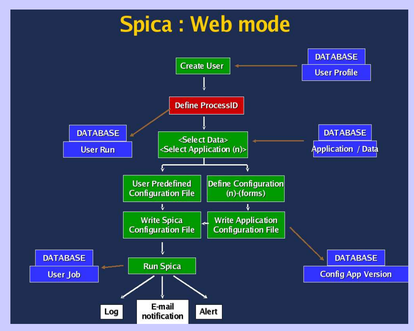
To reduce MegaCam data (the world biggest CCD camera : one square degree), we have designed a pipeline software named Spica (Software Pipeline for Images and
Catalogs Analysis). It produces calibrated images (astrometry, photometry and coaddition/resampling, catalog extraction). It can be installed as a standalone application or as a client/server one.
Once the data are stored on the cluster, Spica checks their integrity and loads them in the SQL database (see RDBix). MegaCam images are uncompressed if needed and sorted out by filter and runid.
The quality of each image is controlled by Qualityfits assessment (see companion poster) and this information is used to select which data could be used by Spica.
There are two ways to use it :

Automated web mode

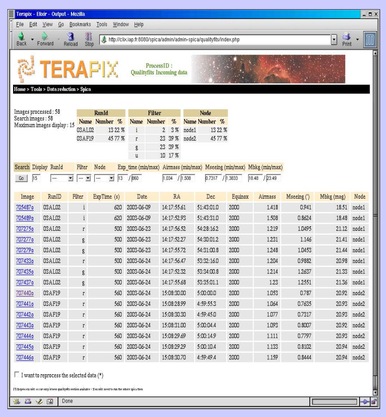


On each node, Perl scripts manage input/output from each pipeline application, database storage and send command line scripts to the spica daemon which run processing according to priority rules. Generated files are sent to the output storage disk. Input data are processed with rules written in configuration files.

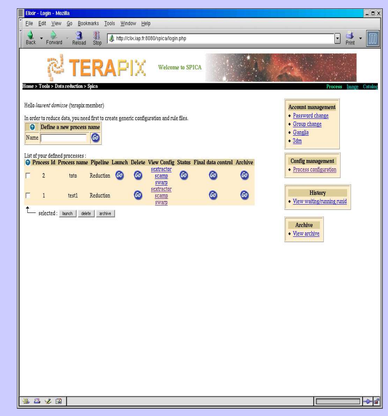
Interactif web mode



A web interface allows authorised users to reduce their own data. They can choose :
1) which data to process according to their preferences,
2) which applications (including configurations) to run.
All steps are recorded in the database for further analysis.



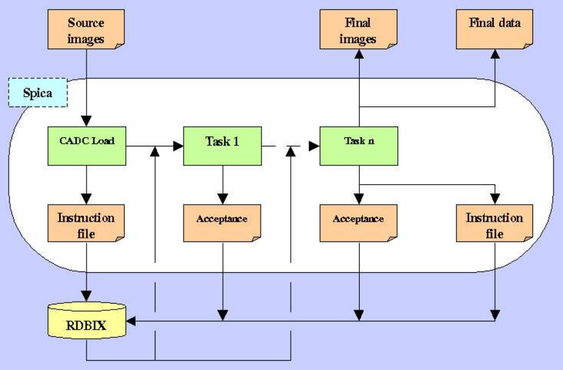
Administration interface written in PHP can check the pipeline status. It can also display spica produced data.



The users have their personal web spaces allowing to create a customised environment. Forms help users to build their setup processes.

RDBix

RDBIX is the memory of the whole pipeline. It is a MySQL relational database designed for the storage of catalogues and metadata produced by the processing software. RDBIX organizes these metadata to maintain history of the images processing.



Each step of the pipeline process communicates with RDBIX to get data about images (coordinates, storage path, ...) and to inform RDBIX, via a piece of software called dbclient, about images modifications. An XML file, called instruction file, serves as a support for this communication with RDBIX. Other data, such as acceptance files (quality data) are also integrated into RDBIX.

RDBIX is based on a completely dynamic design, which allows the database to grow without heavily impacting the performances.
Finally, RDBIX has been designed to run in a distributed environment : each node of the cluster runs a client version of dbClient (and its own pipeline package), while the master node runs a server version of dbClient and hosts the MySQL RDBIX database which is used by every node.

Results

Typical processing of MegaCam images.

On the left is a raw U band 860 seconds exposure. On the right, a stacked U+G+R image composed of 45 Megacam images obtained with Swarp software.

Processing time is around 40 hours on a dual XP1800+ computer (will be down to 30 hours with Opteron processor).

Color image generated by the stiff too which is able also to convert image from FITS format to PNG.

Part of a U band raw image

Same area after image processing

TERAPIX Team