Pipeline Data Processing

Robert Greimel (ING), Jim R. Lewis (IoA) and Nicholas A. Walton (ING)

In the era of 8-m telescopes, there is increasing pressure to operate 4-m class telescopes in the most economic, efficient and effective manner possible. The ING is addressing these three 'E's, in part with its implementation of an improved streamlined data flow system encompassing data acquisition, pipeline processing and subsequent archiving and distribution of raw and processed data products.

The advent of large format CCD arrays and large infra-red detectors has led to an explosion in data volumes. For example, the current data rates at the ING are determined by the Wide Field Camera (WFC) (see e.g. Walton et al., this issue) on the INT which typically generates ~8 GB/night and the IR camera INGRID on the WHT generating some ~4 GB/night. In total, data flows can amount to 15–20 GB/night from all telescopes.

The current day availability of affordable processing power and storage capability has opened the possibility to provide (semi)-processed data products at the point of data origin to the visiting astronomer. These reduced data products will form the core data resource of new 'Virtual Observatories' (see e.g. http://www.astro.caltech.edu/nvoconf/).

1 Data Processing Pipeline

Details of the ING's data processing pipeline algorithms, as implemented for the reduction of imaging data, are described elsewhere (Irwin & Lewis, 2001). The basic pipeline consists of the following steps: linearity correction, bias subtraction, flat fielding and the application of a basic astrometric solution. Additional steps in the pipeline are de-fringing, an accurate astrometric solution, object detection, classification and catalogue generation.

The pipeline can be run in two modes: quick look and science. The goal of the quick look pipeline is to deliver a processed image to the observer within five minutes of image acquisition. This enables immediate assessment of the image quality and instrument performance. The quick look pipeline differs from the science pipeline in two areas: it applies calibration files from the most recent science pipeline run instead of the calibration frames from the current run, and it implements a subset of the full science pipeline reduction, terminating with the de-fringing stage.

The science pipeline provides the observer with reduced data shortly after the end of the observing run. To provide the highest quality data a limited amount of human intervention is necessary, mainly in rejecting poor calibration frames. This pipeline has been running on a Sun© UltraSparc system servicing ING Wide Field Survey Data since August 1998 (Lewis et al., 1999).

2 Powering the Pipeline: Gigawulf

In order to provide an economic processing unit for the pipeline, it was decided to use commodity PC components. Recent advances in the clustering of PC's, making use of the Linux (see e.g. http://www.linux.com) operating system have made this feasible. Linux based PC farm's have been found by a number of groups to offer a powerful and cost effective solution for large computational problems. Indeed, a small number of PC systems have been developed for use in astronomical data processing environments (see e.g. Gravitor http://obswww.unige.ch/~pfennige/gravitor/gravitor_e.html).

The ING data pipeline is a coarse grained parallel processing case. To a first approximation, each science data frame is processed in an identical fashion, with no cross reference to any other. Therefore a night's data can be equally distributed between the nodes. PC clusters are ideally suited for this case (see discussion by Brown, 1999, http://www.phy.duke.edu/brahma/beowulf_advanced.ps).

Gigawulf is a 'Beowulf' type cluster (http://www.beowulf.org for a definition and related links) of eight high end PC's. Each node consists of an AMD Athlon 950 MHz processor with 256 MB of main memory. The seven slave nodes have 30 GB EIDE hard disks while one node, subsequently called the head node, has two 75 GB EIDE hard disks. The head node also has a DDS-3 DAT tape robot as well as a second network card which provides the connection to the telescopes and data archives. The network in the cluster is 100 Mbps apart from the head node, which has a Gigabit connection. A schematic view of the system is shown in Figure 1.

To minimise the operational and maintenance overheads, the Scyld Beowulf extension (Scyld Computing Corporation™, http://www.scyld.com) to Linux (currently based on RedHat's 6.2 distribution, http://www.redhat.com) has been used as the operating system for Gigawulf.

Scyld Beowulf supports standard Linux interfaces and tools. It enhances the Linux kernel with features (provided by bproc, http://www.beowulf.org/software/bproc.html) that allow users to start, observe, and control processes on cluster nodes from the cluster's head node. With this arrangement, software only needs to be configured on the head node. The result is that the cluster appears more

like a shared memory multi-processor computer to a user or developer. This reduces the cost of cluster application development, testing, training, and administration.

3 Pipeline Data Flow

The existing pipeline software has been ported to run on Gigawulf with only small modifications. Scripts have been developed to handle the data flow between the telescopes, Gigawulf and archiving media (DVD-R and DDS-3 tape).

During every night run the quick look pipeline is executed. A process running on Gigawulf automatically looks for newly acquired data files. If a new file is found it is copied to the head node. If the file is a calibration frame it is passed on to a slave node. The important point to note in this step is that the same type of calibration image is always transferred to the same slave node. For example, bias frames always go to node 0, B band flat field images to node 1, and so on. More interesting is what happens if the frame is a target frame. The frame is run through the basic pipeline on the head node using pre existing calibration files from the last science run. For filters in which fringing on the CCD occurs the de-fringing algorithm is also run. The fringe frames used are from the last science run as well. Once the image has been processed it is then copied back to the data reduction computer at the telescope where it can be examined by the observer. The original file is then copied to the slave nodes in a round robin fashion. This assures that we have load balancing between the nodes for processing the data with the science pipeline.

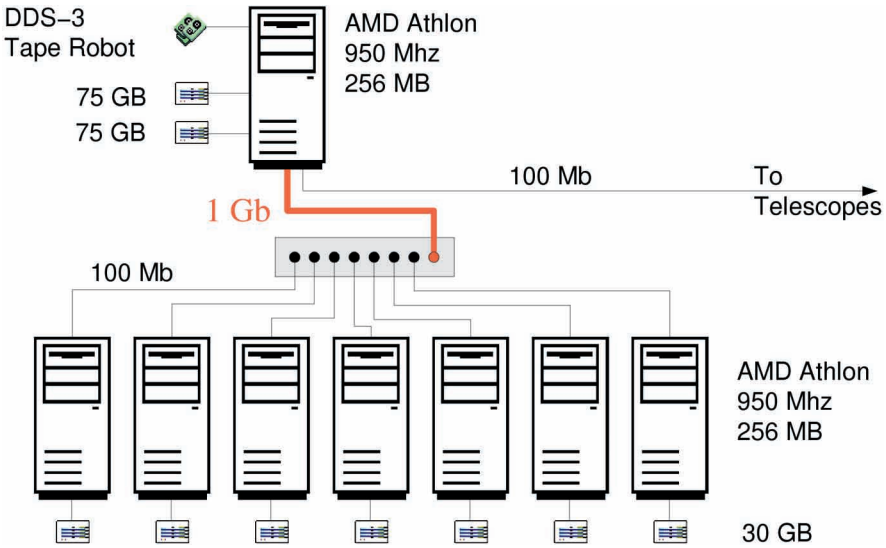

Figure 1. Schematic view of "Gigawulf", a Beowulf cluster consisting of 8 identical high end PCs connected by a high speed network.

At the end of each observing run the calibration images are inspected for quality and bad frames are removed from further processing. The master bias frame is then assembled on node 0. Once finished, the master bias frame is copied back to the head node, which in turn distributes it to all the other slave nodes. Next the master flat field frames are being calculated. This happens mostly in parallel, as the data files for different filters are located on different nodes. When a slave node finished combining its flat field it sends the result back to the head node, which distributes it among the other slave nodes again. Once all the flat fields have been calculated the full pipeline is run on the target frames. This is a completely parallel process. When all

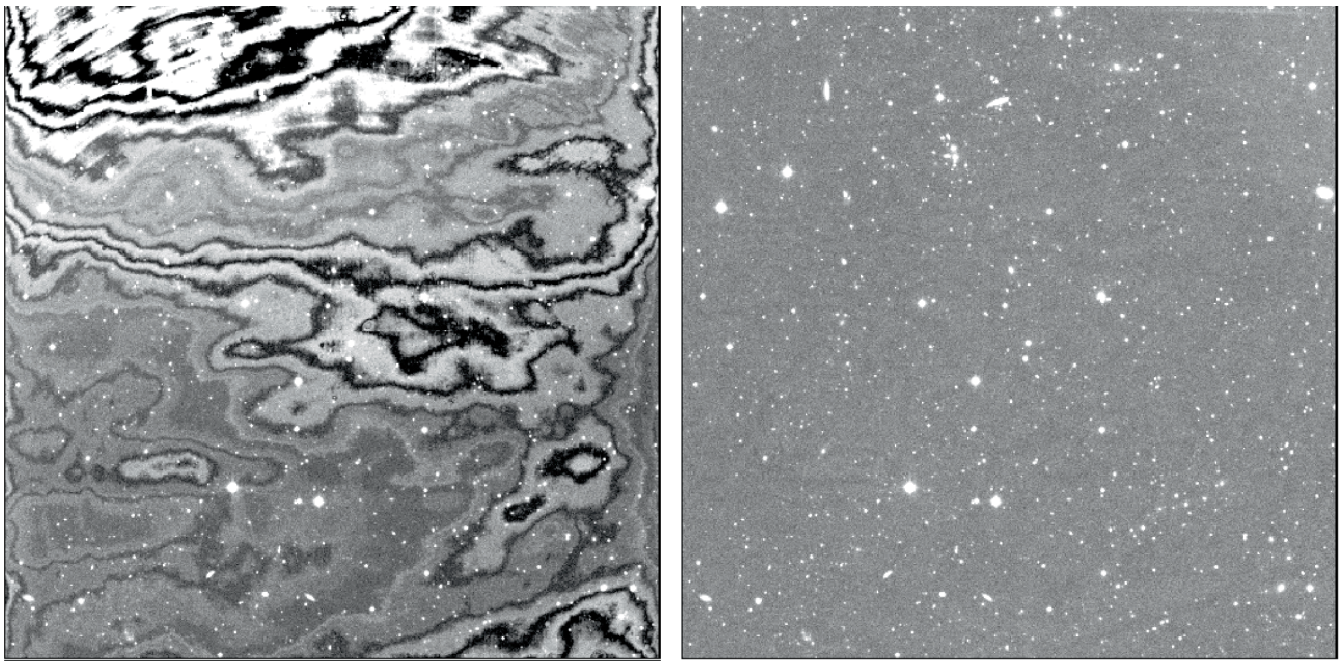

Figure 2. Section of a WFC image before (left) and after (right) de-fringing using Gigawulf.

the nodes have finished processing the data it is then copied back to the head node. The head node writes the data files to DDS-3 tape which are made available to the observer.

Due to the large amount of raw data that is produced by the instruments and the limited data storage capacity on Gigawulf, data cannot be left around after processing for a long time. The raw and reduced data is removed from the system while the calibration data files and object catalogues are transferred to the CD/DVD towers for ingestion into the data and engineering archive (see news article by Don Carlos, this issue).

4 Current Status and Future Plans

Currently, the pipeline (quick look and science) is implemented for the reduction of optical imaging data from the WFC and the WHT Prime Focus Camera (PFC). Operation of the quick look and science pipeline processing on Gigawulf was started in February for both cameras. A sample result of the pipeline reduction process can be seen in Figure 2. The image catalogues will also be used for routine quality control enabling for example monitoring of image quality and throughput in near real time.

The imaging pipeline will be extended to handle data from single CCD cameras (WHT Aux Port and JKT) once these systems have been switched over to use UltraDAS (Rixon et al., 2000) which will be done by summer 2001. A pipeline for near infra-red imaging data produced by INGRID is currently being developed and will be deployed in the very near future. A pipeline will also be introduced for Echelle and multi-fibre spectroscopic data by the end of 2001.

A second Beowulf cluster is currently being acquired. One system will then be exclusively used for the WFC pipeline while the other cluster will run the pipelines of the WHT and JKT instruments. This split is basically determined by the amount of data produced by the different instruments available at the ING.

References:

Irwin, M. J., & Lewis, J. R., 2001, *NewAR* (in press).

Lewis, J. R., Bunclark, P. S., & Walton, N. A., 1999, in Proceedings of ADASS VIII Conference, *ASP Conf Series*, **172**, 179.

Rixon, G. T., Walton, N. A., Armstrong, D. B., & Woodhouse, G., 2000, *Proc SPIE*, **4009**, 132.

Walton, N. A., Lennon, D. J., Greimel, R., Irwin, M. J., Lewis, J. R., Rixon, G. T., 2001, *ING Newsl*, **4** (this issue). ¤

Robert Greimel (greimel@ing.iac.es)

The Second Round of Wide Field Survey Observations

René Rutten (ING)

In July of last year a second call for proposals was sent out, prompting the community in the UK and The Netherlands to submit proposals for survey observations with the Wide Field Camera on the INT. This announcement followed a decision by the ING Board to continue the scheme of survey observations and make available approximately 5 weeks per semester for this work. Survey observations are seen as one of the key roles for the INT and this second call for proposals intends to promote this trend.

The assessment of the proposals this time was left largely in the hands of the UK and NL time allocation panels, who were for the occasion strengthened with two independent assessors. This, it was hoped, would answer the criticism heard in the community during the previous round in 1998. Progress on these proposals will be reviewed annually, based upon which further allocations will be granted.

As for the previous round, the data will be accessible to the wider astronomical community in the UK and the NL though the ING archive based in Cambridge. There will be no propietary period in order to promote fast and wide exploitation of the survey data.

From the proposals received, the following six were selected and will be granted observing time:

– The *Oxford Deep WFC Survey*. PI: Dalton (Oxford).

– *Multi-Coloured Large Area Survey of the Virgo Cluster*. PI: Davies (Cardiff).

– The *Faint Sky Variability Survey II*. PI: van den Heuvel (Amsterdam).

– The *INT Wide Angle Survey*. PI: McMahon (Cambridge).

– The *Local Group Census*. PI: Walton (ING, La Palma).

– An *Imaging Programme for the XMM-Newton Serendipitous X-ray Sky Surveys*. PI: Watson (Leicester).

A brief description of each proposal follows:

The *Oxford Deep WFC Survey* also is a continuation of a previous survey proposal. Its principal scientific aim is to study weak lensing by large-scale structure, the angular clustering of faint galaxies, the clustering of Lyman-break galaxies at high redshift and to measure the luminosity function of faint and distant QSOs.

The project *Multi-Coloured Large Area Survey of the Virgo Cluster* will complete the survey that has already