ESO's Next Generation Archive System

A. WICENEC, J. KNUDSTRUP, S. JOHNSTON

Abstract

Early in July 2001 ESO/DMD installed prototype versions of the archiving and buffering units of the Next Generation Archive System (NGAS) at the 2.2-m telescope in La Silla. The two units are the on-site part of an archive system we are currently testing for high data rate/ high data volume instruments like the Wide Field Imager which is mounted at the 2.2-m telescope. The NGAS concept is built around two ideas: the use of cheap magnetic ATA-100 disks as the archiving media and a highly flexible and modular software called NG/AMS, NG Archive Management System. The main goals of the whole system are scalability in terms of data volume, but also the ability to support bulk data processing either by fast data retrieval or by opening the computing power of the archive for data-reduction close to the data themselves. In fact the NGAS scales in such a way that it is possible to process all the data in the archive within an almost *constant* time. In this article we present an overview of the NGAS concept, the NGAS prototype implementation and some of the experience we have made during the first month of operating the system in a real observational environment. We also present the infrastructure of the main archive which supports scalable, decentralised processing, both of which are essential for large-scale scientific programmes submitted to a Virtual Observatory.

Figure 1.

1. Introduction

With the advent of wide-field mosaicing CCD cameras, the data rate of several observatories around the world is literally exploding. While some of these instruments are already in use (e.g. WFI@2p2; CFHT 12k; SLOAN) other, even bigger ones, are under construction or planned (Omegacam, Megacam, VISTA). The current archive system at ESO consists of DVDRs which are written in two copies at the observatory sites. One of the copies is sent to the ESO headquarters in Garching to be inserted into a DVD jukebox. In this way the data are quasi on-line in a juke-box about 10 days (mean) after the observations have been carried out. Given the current set-up of the system, it is possible to archive up to about 15 to 20 GB per night. The data rate coming from the ESO Wide Field Imager (WFI@2p2) easily hits this limit in a typical night and is a lot higher for exceptional programmes (up to 55 GB/night). The expected data rates of Omegacam and VISTA are 4 times and 8 times higher than the one from WFI@2p2, respectively. In order to be able to cope with such data rates, ESO initiated a project (Next Generation Archive System Technologies, NGAST) to come up with alternative archive solutions.

Figure 2: This figure shows three of the NGAS units mounted in a rack. They are fully equipped with eight disks each, giving a total of 1.65 terabyte of on-line storage capacity with three processors running at 1.2 GHz each.

2. Requirements

A new archiving system must resemble the current costs and operational scheme as closely as possible. For the costs it is clear that one has to account for the pure hardware costs, the operational costs and the maintenance costs. The hardware includes the costs for the consumable media, readers, writers (if any) and computers. Apart from scalability in terms of data volume and throughput at the observatory, a Next Generation Archive System has to fulfil a number of basic additional requirements in order to be able to cope with future challenges:

- Homogeneous front-end (archiving at observatory) and back-end (science archive) design
- Access to archive shall be scalable, i.e. the number of entries and volume of data shall not affect the access time to single data sets
- Support bulk data processing mainly for the quality control process, but having Virtual Observatory projects in mind
- Processing capabilities shall scale along with archived data volume, i.e. it should be possible to process all data contained in the archive
- Economic solution using commodity parts to reduce overall costs

The main goal of the first point is to limit maintenance costs, operational overheads and the time-to-archive. Time-to-archive is the total time the complete system needs until data is on-line and retrievable (disregarding access restrictions) from the science archive. The support for bulk data processing is mainly driven by the fact that ESO is already now processing about 50% to 75% of all data, in order to ensure the data quality for service mode programmes, monitor the telescope/instrument parameters and provide master calibrations frames for the calibration data base. With the very high data rate instruments the demands for processing and storage capabilities for quality control will grow tremendously.

3. ESO's Prototype Solution

For the implementation of the prototype units we chose a particular hard-

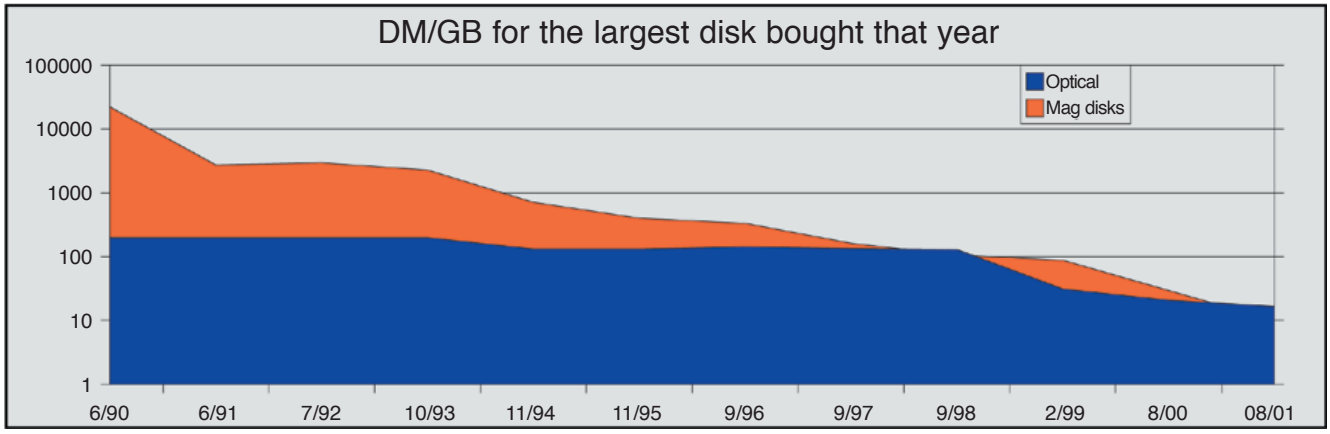

Figure 3: This graph shows the price per gigabyte for the biggest disk bought by ESO during the year indicated on the x-axis. Most of the disks are SCSI disks; only during the last two years also IDE disks have been included in the graph.

ware and software configuration and made some implementation decisions.

3.1 Hardware configuration

The main hardware implementation decisions include the following points:

- Magnetic disks with ATA-100 interface in JBOD configuration
- Standard PC components, like mainboard and network card
- 19 inch rack-mount case with redundant power supply and 8 slots to host the data disks
- Removable hard disk trays (hot-swappable)
- Eight port SCSI to ATA-66 PCI card (3ware Escalade 6800)

3.2 Software configuration

The main software implementation decisions include the following points:

- Linux as the operating system
- Next Generation Archive Management System (NG/AMS) server software written in Python
- Multi-threaded HTTP server implementation
- URL-based command interface to the NG/AMS server
- Plug-in architecture to provide methods for different data types and processing capabilities
- XML-based configuration and message passing

4. NGAS Archive Layout

Magnetic disks are consumables in NGAS and for every eight disks a PC will be added to host those disks and bring them on-line in the archive. Right now we have two operational central units (NCUs, NGAS Central Unit) in Garching with 12 completed disks mounted. Since each NCU is capable of hosting 8 disks, we have to add another NCU end of this year. Each of the NCUs is running a NG/AMS server and every eight NCUs together with a master unit (NMU, NGAS Master Unit) and a network switch will form a Beowulf cluster for processing and data management. In this way CPU power is scaled up together with the data, and processing of all the data in the archive for the first time becomes feasible. The time needed to process *all* the data in an NGAS archive can be kept almost constant as long as the data-to-CPU ration is kept constant. In fact, we are planning to control all the data holding in the NGAS archive permanently by performing CRC checks on a file-by-file level. The NG/AMS is prepared to support large-scale processing in the archive by distributing tasks to the nodes where the data resides. The same software with a slightly different configuration is used to control the archiving process at the telescope as well as the main archive at ESO headquarters in Germany. As soon as a new magnetic disk is formatted and registered in the NGAS data base the location and contents of it is always traceable. Archiving is done in quasi real-time and the data are immediately accessible through the standard archive interface. In principle it is possible to retrieve frames right after they have been observed, but for security and data permission reasons there is a very restricted, fire-wall protected access to the archiving units on the observatory sites. In addition to fire-wall security NG/AMS is configurable to either deny or grant permission to retrieve archived frames.

The ESO Science Archive Facility imposes some additional requirements for the back-end implementation of NGAS. These include seamless integration in the current archive query and request handling system. NGAS has to support fast retrieval of single files, lists of files and files belonging to a specific programme, to name a few retrieval scenarios. Moreover, retrieval of FITS headers only, production of previews and archiving of master calibration frames has to be supported by the main archive part of NGAS. The operational, maintenance, power consumption and physical volume aspects of an ever growing number of PCs have to be addressed as well.

5. Milestones and Performance

The front-end (archiving) system consisting of two NGAS units was installed at the 2.2-m telescope beginning of July 2001. Since then, this prototype installation is archiving data from the ESO Wide Field Imager (WFI). In short, the milestones and performance numbers of this non-optimised prototype look like the following:

- NGAS unit prototype installation on La Silla: July 3 to 13, 2001
- Start of operations on La Silla: July 7, 2001
- First terabyte of data controlled by NGAS: September 18, 2001
- Installation of first two NGAS units for the main archive (NCUs) in ESO HQ: September 25, 2001
- Commissioning and acceptance of front-end NGAS on La Silla: December 2001
- Commissioning and acceptance of back-end NGAS in ESO HQ: February 2002

The prototype front-end NGAS is not yet fully optimised for performance, but the time-to-archive was always shorter than the production time of frames by the instrument. The data flow from WFI between July and September 2001 was 13.7 GB/night (median) with a maximum of 53.8 GB in a single night. The overall throughput of the archiving process during the same period was 3.17 MB/second, including compression and replication of the files. The hardware used in the NGAS units provides very fast write access to all the eight data disks in parallel, summing up to about 100 MB/second (measured), thus there is plenty of room for improvement of the overall system performance.

6. Overall Costs

The overall hardware costs of the NGAS have been carefully calculated.

Media	DM/GB in Juke Box	Number of Media/Terabyte
CD-R	70.3	1625.4
DVD-R (3.95 GB)	16.0	276.8
DVD-R(4.7 GB)	11.6	227.6
DVD-RAM (2 × 4.7 GB)	10.8	113.8
MO	36.8	204.8
SCSI disk farm	8.8	5.7
Sony 12″ Optical Disk	215.7	160.0
RAIDZone NAS	17.7	14.2
NGAS	8.7	14.3

Table 1: Comparison between different media in terms of price/GB of storage and number of media/Terabyte. The SCSI price is remarkably low, because of one of the new 180 GB disks we bought recently under exceptionally good conditions. This kind of comparison is one of the planning tools for the medium term planning of archive media.

Compared to other on-line or quasi on-line random-access data-storage solutions it is the cheapest solution, providing at the same time very low operational costs and very few storage media. Moreover, it is the only solution providing enough computing power to process all archived data with no additional costs. Especially the operational overhead in terms of manual operation drops quite dramatically in the case of the ESO WFI, from about 2 hours/day with the currently used tape procedure to 20 minutes/week. The time-to-archive is also substantially lower (of the order of seconds), because even compared with the DVD system used for the VLT, the data are only on-line when they arrive at ESO HQ about 10 days after the observations; with the tapes the delay is much longer. For very high data volume instruments the number of media/Terabyte becomes a critical parameter, for the production, management and handling.

7. Future of NGAS

NGAS has proven to be a reliable and fast system. Since NGAS is an operational model on top of a hardware/software system, quite different from the one currently used, an implementation for other ESO telescopes/instruments will still take some time. While the prototype system on La Silla will go operational end of this year, other installations on La Silla and Paranal are not required, because the DVD system is able to deal with the data rate of the currently installed instruments. We are planning to use NGAS first for very high data rate instruments like Omegacam which will be operated on the VLT Survey Telescope beginning in 2003. MIDI (VLTI) is another candidate for a NGAS installation. NGAS will certainly be evaluated as one of the building blocks of the Astrophysical Virtual Observatory in the area of scalable archive technologies and it also is already evaluated in the framework of the ALMA archive.

References and Acknowledgements

Most up-to-date information can be found under http://archive.eso.org/NGAST

ESO Wide Field Imager: http://www.ls.eso.org/lasilla/Telescope/2p2T/E2p2M/WFI

AVO: http://www.eso.org/projects/avo

ALMA: http://www.eso.org/projects/alma

We would like to thank especially the 2.2-m telescope team, the La Silla Archive team and Flavio Gutierrez for their on-going support.

News from the 2p2 Team

Personnel Movements

In September we welcomed new team member Linda Schmidtobreick from Germany. Linda is a new ESO Fellow and will be working primarily with the 2.2-m. Before joining ESO, Linda held a two-year postdoctoral position at Padova, Italy. Her research interests include stellar populations, cataclysmic variables and the structure of our Galaxy.

September also saw us farewell Heath Jones from his La Silla duties. Heath will complete the 3rd year of his ESO fellowship at Cerro Calan.

WFI Images

As a Christmas present from the 2p2 team, we have included here colour images of the Dumbbell and Triffid planetary nebulae. These were created from only one of the chips of the Wide Field mosaic by team member Emmanuel Galliano. The separate B, V and R images used to make the pictures were taken under average seeing conditions on June 11/12, 2001.

Figure 1: The Dumbbell nebula from 10 minutes B, V and R images.