ESO's Next Generation Archive System in Full Operation

Andreas Wicenec, Jens Knudstrup
(ESO)

Considerations of the technical feasibility and the cost implications of a disk-based archiving system to store digital observations coming from the ever growing suite of ESO telescopes and instruments began in 2000. The so-called Next Generation Archiving System (NGAS) started archiving data in a prototype system in 2001. Now the second generation of NGAS hardware has been installed in the new ESO data centre and about 98 % of all data since 1998 have been migrated onto disks hosted on NGAS computers. In addition all data currently produced by ESO instruments is directly archived onto NGAS hosts both in La Silla and Paranal. Currently the ESO archive keeps about 125 TB of data online and the system has been scaled up to cope with the next data wave coming from VISTA and OmegaCAM.

The NGAS concept

The release of a white paper (Wicenec and Pirenne 2000) describing the technical feasibility and the cost implications of a disk-based archiving system marked the start of a new chapter in a quite different area than commonly described in the ESO Messenger. This story is about persistently storing digital observations coming from the ever growing suite of ESO telescopes and instruments. The so-called Next Generation Archiving System (NGAS) started out as an idea and a feasibility study. In the first Messenger article from December 2001 (Wicenec et al. 2001), it was still described as a prototype system. Early in July 2001 the Data Management Division (now Data Management and Operations Division) installed prototype versions of the archiving and buffering units of NGAS in the control room of the 2.2-m telescope in La Silla. The two units were the on-site part of an archiving system we were testing at that time for high data rate/high data volume instruments like the Wide Field Imager mounted at the 2.2-m telescope (WFI@2p2). The original NGAS concept was built around two basic ideas: use of cheap commodity hardware with magnetic ATA-100 disks as the archiving and transport media; a highly flexible and modular software called NG/AMS, Next Generation/Archive Management System. The main goals of the whole system are scalability and the ability to process bulk data within the archive itself. In fact NGAS scales in a way so that it is possible to process all the data in the archive within an almost constant time. In the meantime technology advances have led to the usage of SATA2 rather than ATA-100 disks, but that is quite a minor detail. On the technology side we had to change hardware components for newly procured NGAS computers several times, but the first computers installed at La Silla were only replaced and upgraded in 2005 in order to provide more redundancy and to be able to capture all data from all the instruments operating at La Silla. At the same time the NGAS systems were moved to the RITZ (Remote Integrated Telescope Zentrum).

Figure 1: Close-up view of some of the NGAS machines of the primary archive cluster in the new ESO data centre.

On the Garching side, in the main archive, we started with eight computers, with eight disk slots each. The disks used in 2001 had 80 GB and were close to the optimal price/capacity ratio at that time; they were filled up with one week of typical WFI@2p2 operation and were then ready to be shipped to Garching via the diplo bag. The latter highlights another key point of the NGAS concept, the shipping of hard disks and the full traceability of the shipment procedure. In NGAS operational terms, magnetic disks are consumables and data can be migrated freely from one disk to another. In fact every file stored on NGAS is fully virtualised in the sense that access to a file is solely controlled using a unique NGAS file ID. There is no need to know the actual computer, disk or directory path where the file is located. A request for a file with a given ID can be issued to any NGAS computer available on the network; the NG/AMS will figure out where the closest available copy is available and deliver that copy to the requester.

The last feature of the NGAS concept is the interoperability of NG/AMS with other components of the Data Flow System (DFS) and other client software or direct human users. In order to minimise the implementation impact on both the NG/AMS server and the clients, it was decided very early on in the project to use an existing very simple protocol which is as widely available as possible. Consequently NG/AMS is actually implemented as an HTTP server. All available commands can be issued through standard HTTP clients, including web browsers. NG/AMS is supposed to be used through software rather than directly by humans, and the latter is not recommended because the core NG/AMS does not provide a real page-oriented web interface.

NGAS requirements

Back in 2001, a new archiving system had to resemble the operational scheme of the existing system as closely as possible and be similar in terms of cost. For the costs, it is clear that one has to account for the pure hardware costs, as well as the operational and maintenance costs. The hardware includes the costs for the consumable media, readers, writers (if any) and computers. In order to be able to use magnetic disks as an archiving

media, the overall system has to fulfil a number of basic requirements:
– Homogeneous front-end (archiving at observatory) and back-end (science archive) design;
– Access to archive 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;
– Processing capabilities should scale along with archived data volume, i.e. it should be possible to process all data contained in the archive;
– An economical solution using commodity parts to reduce overall costs. This consideration also includes power economy, whereby unused servers are switched down after a configurable idle time and then woken up for a request, in order to save power;
– Possibility to use the magnetic disks as a transport medium for the bulk data.

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 the data is online 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 almost 100 % of all data, in order to ensure the data quality for service mode programmes, monitor the telescope/instrument parameters and provide master calibration frames for the calibration database.

NGAS archive facts and facets

The currently installed NGAS cluster for the primary archive can host up to 150 TB of data, distributed across 24 machines with 24 disks each (see Figure 1). In this configuration it is already prepared to start receiving and storing VIRCAM and OmegaCAM data from VISTA and VST respectively, in addition to the data stream from the VLT, VLTI, and La Silla telescopes. As of end July 2007, the primary archive holds more than 7.2 million individual frames obtained by ESO instruments (in general one observed frame results in one file in NGAS). The total number of files stored on NGAS at present amounts to almost 30 million. The large number is due to the fact that at the moment at least two copies of each file are kept in the primary archive, because the secondary archive is still in the process of being populated before becoming fully operational. In addition also the master calibration frames produced as a part of the quality control process of most of the ESO instruments, as well as auxiliary files and log files, are archived on NGAS. The archive system design and the use of commodity hardware in both the primary and the secondary archive meet the goal of keeping development, maintenance and operations costs low.

With the new VO-compliant science archive interfaces to be released by the end of this year, we expect that significantly more scientists will be able to exploit and use the archived data beyond its original scientific intent. About 1 million requests every month are served by the NGAS archive main servers. Most of these requests are internal to ESO operations, quality control processing and archive maintenance requests, but without the online nature of the NGAS archive all these processes would need substantially more time and staging disk space. The new VO compliant interfaces require also direct access to the data, essentially through web links (URLs). This introduces a new paradigm for access to ESO data by external users, because up to now the data is only served in an asynchronous way and requesters receive e-mails upon completion of their requests. In the future the data will be almost directly served by NGAS to the global astronomical community. In order to be able to cope with these new requirements, the network infrastructure of the NGAS cluster will be changed as well, in order to make full use of the intrinsic parallelism.

On the front-end side NGAS is archiving between 1000 and 6 000 new observations every night. This rate is mainly dependent on the number and type of instruments operated on the mountains and on the weather conditions. Some instrument modes are quite demanding for the rate of archiving files, which may in some exceptional cases rise to several hundreds of thousands of files per day. In addition to the La Silla Paranal instruments, the raw data of the WFCAM instrument on the UK Infrared Telescope (UKIRT) in Hawaii and the data from the APEX sub-millimeter telescope are also archived on NGAS. In particular, the data from WFCAM poses quite an additional load and required a special set-up, because, firstly, this instrument produces about 200 GB of data every clear night and, secondly, the data is archived through the network from Cambridge, UK.

The NGAS implementation

As already mentioned NGAS is an integrated hardware and software solution for bulk archiving, archive management and basic large-scale file processing. This means that both the hardware and software configurations have to be kept under strict configuration control. For the hardware this includes not only the single computers, but also the cluster and network configuration, the racks, the cooling concept, the connection to external systems, like the quality control processing cluster and the secondary archive, and the compatibility between the front-end mountain-top systems and the primary archive.

The hardware

As an integrated system where magnetic disks are to be used as consumables, we had to be more strict in the selection and maintenance of the machines and their components. In particular the requirement to use the disks as a data transport medium requires that the machines on the mountains and the machines in Garching use compatible disk slots and disk trays. Since removing and mounting disks is a very regular and standard procedure, the mechanics have to operate both smoothly and reliably and the parts have to be rigid enough to perform many mount/remove cycles. At the same time the trays should be compact, provide efficient cooling and not be too expensive. The disks are shipped in their trays and are then mounted in a different machine, thus the slots have to be fully compatible. In order to have better control on such details we have chosen a computer selection process which involves the specification of computer parts rather than a model. All of the parts are commodity parts and thus easily

available from many vendors. The new ESO data centre, which was inaugurated on 27 July, hosts the 24 machine NGAS cluster (see Figure 2). With a future upgrade of this cluster with high capacity disks and a different RAID configuration, these 24 machines can host up to 0.5 Petabyte of data.

The software

The machines are installed using a standard Scientific Linux OS installation with customised packages, and some customised system configuration, followed by an installation of the NG/AMS software. NG/AMS is written in 100 % pure Python. Python is an object-oriented scripting language in wide use; for more information on Python see *http://www.python.org*. The experience of writing and maintaining a rather big server application in Python is quite positive, mainly because of the clarity and compactness of the language. Things like 'Segmentation fault' and the related, sometimes tedious, debugging sessions simply do not occur and the very clear object orientation of the language allows for a clean and proper design of even complex object relations. The high-level built-in modules add to a very efficient way of programming, resulting also in a comparably low number of code lines.

NG/AMS in its core is a multi-threaded HTTP server. The software implements 20 custom commands, where the most important ones certainly are ARCHIVE and RETRIEVE. These commands are accessible through standard URLs of the form: *http://ngasserver:7777/RETRIEVE?file_id=this_is_my_file*[1].

The ARCHIVE command supports both push and pull of data, i.e. a file can be pushed by a client using a standard HTTP POST method or it can be pulled from some other server, for instance an ftp server, if a URL is specified. NG/AMS features full virtualisation, which means that on retrieval the only thing one has to know is the NGAS file-id and the name and port of *one* accessible NGAS server of the archive in order to access any file anywhere on any NGAS machine of the ESO archive. Thus, in principle, one could retrieve the last frame observed in Paranal by just knowing its file-id. In practice, however, this is not possible because access restrictions apply for computers running in the La Silla Paranal Observatory. To ensure the safety and security of the data, none of the operational NGAS servers are accessible from outside ESO and even inside ESO only a few machines and users can access them.

In order to trace the location of every file in the system, NGAS uses a database which contains information about all NGAS hosts, disks and files. There are additional tables to control the history and the location of disks even if they are not on-line, have been retired or are currently travelling to one of the sites. By default NGAS always keeps two copies of every file and this is rigidly controlled and checked throughout the operations. Without some effort and special permission it is not possible to remove a file if that would result in less than two copies being available, and, anyway, being an archiving system, deleting files is a protected action. The consistency of the contents of the database and the files on disks is checked periodically; this includes the calculation of a checksum for every file. If discrepancies are detected the software sends an notification e-mail to the archive administrators.

There are a number of advanced features in NGAS, which are currently not, or only marginally, used in the ESO operational environment. These include a power saving mode, where NGAS nodes will go into sleep mode and are automatically activated to handle requests and for consistency checking. Another as yet unimplemented feature is the processing of files upon request and the return of only the result of this process.

NG/AMS is very flexible and configurable. It provides 10 different plug-in types to customise the behaviour of various parts of the system, these include plug-ins which are executed when the system goes online and offline, when data is archived or registered, and for processing and checksum calculation. For the power-saving mode there is a plug-in which causes the node to go to sleep and another one which wakes it up. NG/AMS also provides a subscription service, where one node can subscribe for data being newly archived on another node.

Figure 2: The current 24-machine NGAS cluster in its new location, the ESO data centre. Each of the three racks weighs about 1 000 kg and uses a sophisticated dual cooling system, located behind the narrow grey doors. The cluster can host up to 576 SATA disks. With the currently used 400 GB disks this translates to a total capacity of 156 TB in 48 RAID5 disk arrays.

[1] This is a fake URL.

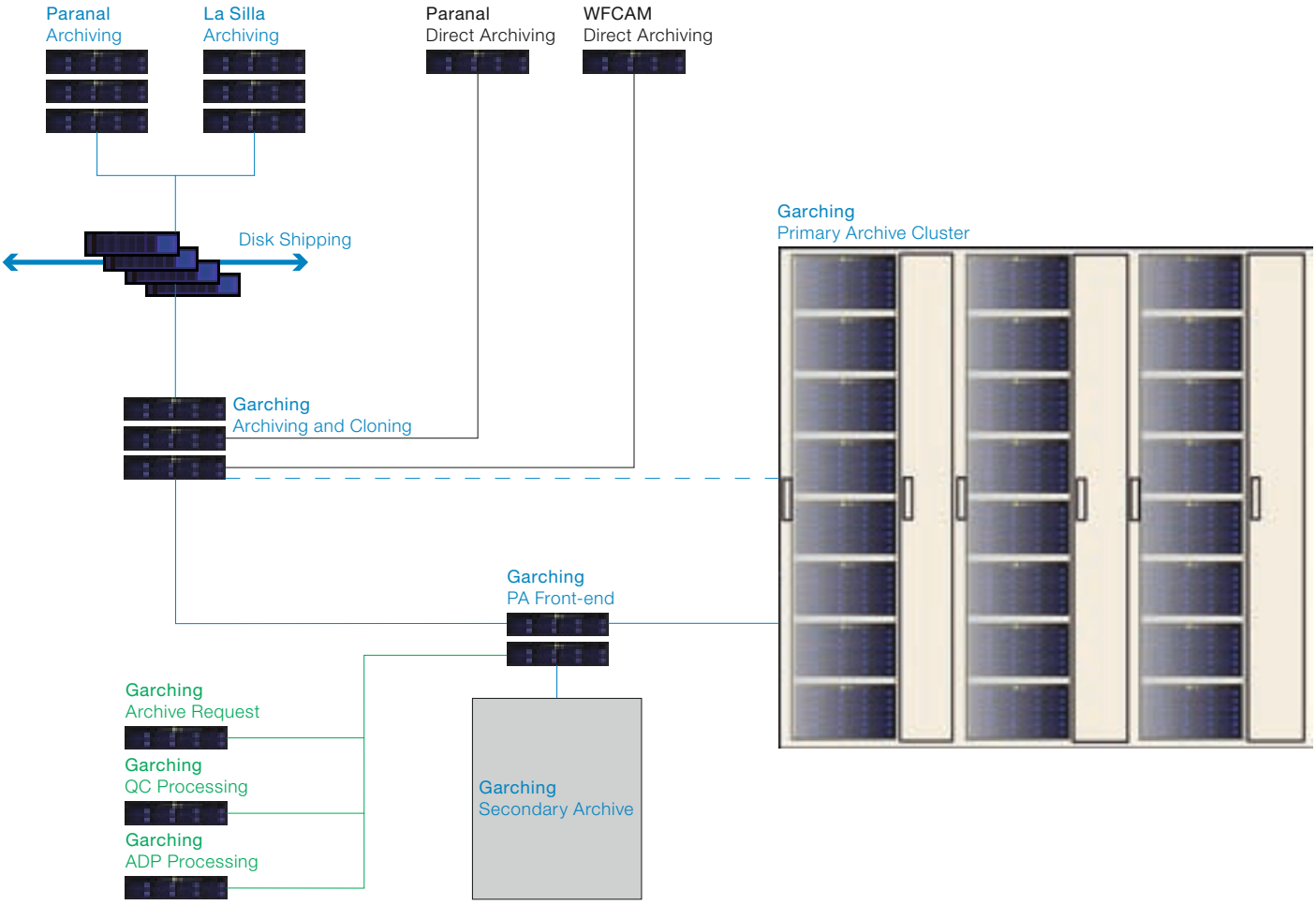

Figure 3: Schematic view of the complete NGAS data flow. The main data from the La Silla Paranal observatory is archived on NGAS machines in the observatory and then shipped to Garching. This part of the observatory data flow is highlighted with blue letters and lines. The green letters and lines highlight data flowing from and to the NGAS primary archive cluster, to post observation requests, quality control and advanced data product processing. The black letters and lines mark custom configurations for pre-imaging data from Paranal and for the data from the UK WFCAM camera.

For this subscription mechanism there is a filter plug-in to be able to subscribe only to data which fulfils certain criteria. In addition the latest version of the ALMA branch of the software also provides a mechanism to register and execute new commands. This flexibility enables the usage of NG/AMS in many different situations and hardware set-ups. The core software is completely independent of the hardware and can be run even on a laptop.

NGAS operations

As can be seen in Figure 3 NGAS, is operated at three different ESO sites. Both observatory sites have a small three-machine cluster, where only one machine is actually used for archiving data, one is a disk handling unit and one is a spare. The direct archiving from Paranal and Cambridge, UK, is done using a stand-alone NGAS archiving client and is thus not a full NGAS installation. The Garching installation is a bit more complex with the complete primary archive and a number of archiving and disk handling units. There are several external applications, which use NGAS to archive or request data. These include the standard archive requests, where the ESO request handler is the application which executes the actual retrieval. The quality control (QC) processing is executed on its own cluster of machines and retrieves almost all observed frames from NGAS using a direct client. After finishing the process, the QC scientists of the Data Flow Operations department also archive the results on NGAS. This involves mainly the master calibration frames, but a new application supporting the archiving and proper registration of the quality control science products is currently being tested. Another heavy user of the NGAS archive is the processing carried out by the Advanced Data Products (ADP) processing of the ESO Virtual Observatory Systems department. Also in this case a large fraction of the whole archive is requested, processed and the results are archived and registered.

The secondary archive essentially is a back-up of the whole NGAS cluster on a Sony Petabyte tape library. A special application has been developed to interface NGAS with the commercial application, called ProTrieve, controlling the

tape library. Obviously the tape library has to be able to store the same amount of data as the NGAS cluster and thus some care has been taken to procure a scalable solution for this as well. All the hardware for the primary archive and the secondary archive, as well as ProTrieve, have been procured and are maintained through a frame contract with the Munich-based company Kayser-Threde.

NGAS activities elsewhere

NGAS is not only used by the ESO archive, but has also been chosen by ALMA (*http://www.eso.org/projects/alma*); it hosts almost the complete ST-ECF Hubble Space Telescope Archive (*http://www.stecf.org/*) with about 10 TB of data; and it is used for the long-term VLA archive at the NRAO in Socorro. It is also under investigation for the Hubble Legacy Archive (HLA) activities (see *http://hla.stecf.org* for details) both at ST-ECF and at STScI in Baltimore.

Milestones and performance

The front-end system consisting of two NGAS units was installed at the ESO 2.2-m telescope in the beginning of July 2001. Since then, this prototype installation has evolved into a rather big operational system, which is now archiving and controlling practically all data collected by ESO instruments. The historical timeline is:

- NGAS unit prototype installation La Silla, 3 to 13 July 2001
- Start of operations on La Silla, 7 July 2001
- First terabyte of data controlled by NGAS, 18 September 2001
- Installation of first two NGAS units for the main archive at ESO Headquarters, 25 September 2001
- Commissioning and acceptance of front-end NGAS on La Silla, December 2001
- Commissioning and acceptance of back-end NGAS at ESO Headquarters, February 2002
- Installation of the VLTI NGAS on Paranal, January 2004
- Upgrade of the NGAS installation in La Silla, January 2005
- Upgrade of the NGAS installation in Garching and data migration, first half of 2005
- Installation of new NGAS hardware on Paranal and La Silla, supporting the archiving of all data from all instruments, November 2006
- Almost all data migrated from DVDs to NGAS, March 2007

The 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 instruments. The typical throughput of the archiving process on the current hardware is up to 15 MB/second, including compression and replication of the files. The hardware used in the NGAS units provides very fast write access to the data disks in parallel, summing up to about 350 MB/second (measured), thus there is plenty of room for improvement of the overall system performance.

One bottleneck is the database access, which sums up to a non-negleglible load on the database server in Garching, because the whole NGAS system is writing its information into the same database. The Paranal and La Silla NGAS databases are being replicated to the Garching database as well. With 30 million archived files in total and up to many hundreds of thousands of files on a single volume, the queries have to be analysed and optimised in order to improve the transaction times. For safety reasons NGAS performs many consistency checks and holds up to three copies of the data during certain requests, for some type of requests this high safety level might not be necessary and could be lowered. This has been already implemented on the ALMA branch of the NG/AMS for the data retrieval command. The performance increase is about a factor of 25 and clearly shows the potential of such optimisation work.

Future of NGAS

NGAS has proven to be a reliable and fast system. It has managed many tens of millions of files, where the problems were mainly due to hardware/software interactions. Given the off-the-shelf inexpensive hardware used, the reliability in fact is quite remarkable and confirms the study of Schroeder and Gibson (2007) which essentially states that this kind of inexpensive hardware does not fail significantly more often than very expensive hardware. Since now the NGAS archive captures all data from the La Silla Paranal Observatory and almost all 'historic' ESO data has been migrated as well, NGAS really has become *the* persistent storage backend of the ESO archive. The flexibility of the NG/AMS software allows a fast customisation of the full system to other requirements, like ALMA and HST, or adjustments to different hardware. Quite new developments are the usage of NGAS for one of the reference implementations of the VOSpace standard of the International Virtual Observatory Alliance (Graham et al. 2007). As a side-product of this, a prototype WebDAV (Web Distributed Authoring and Versioning is a proposed standard of the Internet Engineering Task Force (IETF), see Dusseault 2007) interface on top of NGAS has been implemented, which exports NGAS as a mountable file system (Harrison et al. 2006).

Acknowledgements

We would like to thank especially the 2.2-m telescope team, Flavio Gutierrez and Jose Parra and the La Silla Paranal Data Handling Administrators (DHA) for their invaluable support during the installation and operational phase of NGAS. In addition we would like to thank the SOS and the EDAT teams and in particular Dieter Suchar for their support in the design, procurement, set-up and operation of the NGAS hardware. We would also like to thank Nathalie Fourniol and the SAO and SEG groups for their suggestions for improvement and many fruitful discussions.

References

More detailed information on NGAS is available at *http://www.eso.org/projetcs/ngas*

Dusseault L. (ed.) 2007, RFC 4918, *http://www.ietf.org/rfc/rfc4918.txt*

Graham M. et al. 2007, VOSpace service specification, Version 1.01, *http://www.ivoa.net/Documents/latest/VOSpace.html*

Harrison P. et al. 2006, in "Astronomical Data Analysis Software and Systems XV", ASP Conference Series 351, 402

Schroeder B. and Gibson G. A. 2007, presented at the 5th USENIX conference, *http://www.cs.cmu.edu/bianca/fast07.pdf*

Wicenec A. and Pirenne B. 2000, The Next Generation of Science Archive Storage, *http://www.eso.org/projects/ngas/ngas-whitepaper.pdf*

Wicenec A., Knudstrup J. and Johnston S. 2001, The Messenger 106, 11