Operating a Petabyte Class Archive at ESO

Dieter Suchar*[a], John S. Lockhart[a], Andrew Burrows[a]
[a]ESO, Karl-Schwarzschild Str. 2, 85748 Garching near Munich, Germany

ABSTRACT

The challenges of setting up and operating a Petabyte Class Archive will be described in terms of computer systems within a complex Data Centre environment. The computer systems, including the ESO Primary and Secondary Archive and the associated computational environments such as relational databases will be explained. This encompasses the entire system project cycle, including the technical specifications, procurement process, equipment installation and all further operational phases. The ESO Data Centre construction and the complexity of managing the environment will be presented. Many factors had to be considered during the construction phase, such as power consumption, targeted cooling and the accumulated load on the building structure to enable the smooth running of a Petabyte class Archive.

Keywords: Petabyte, ESO Primary and Secondary Archive, ESO Data Centre, Compute Cluster, Fast Cache

1. INTRODUCTION

The ESO science archive is a powerful resource for research and operations that stores the observations obtained with the major telescope on La Silla since 1991, as well as the whole record of observations obtained with the VLT on Paranal since the beginning of its operations in 1999. Its holdings have increased by orders of magnitude during these years as new instruments and larger format detectors have become operational. In the coming years, the ESO archive will continue to experience major growth as two new survey telescopes, VISTA and VST, and their large format cameras enter operations producing unprecedented data rates at ESO facilities. The second generation of VLT instruments will also increase substantially the data output of the Observatory. The ability to store, protect, and provide reliable accessibility to these vast data holdings can only be provided by a carefully designed Data Centre based on state-of-the-art technology, which we describe in this article. The ESO Scientific online archive in 1998 consisted primarily of 2 CD jukeboxes hosting approximately 1000 CDs with a total capacity of 0.5 TB. In 2005, the entire ESO science data archive was based on a heterogeneous set of digital media and media management systems, including: off-line CDs and DVDs that must be manually mounted for reading, DVDs mounted in four high capacity jukeboxes and small RAID systems attached to Solaris systems [1]. The main ESO data holdings were stored in a Linux-based hard disk farm consisting of 100x200 GB PATA disks, 25x250 GB SATA disks, and 48x400 GB SATA disk [1]. Over the past 10 years the data volume of the ESO Scientific Archive has been growing to 75 TB [3] located on 576 SATA hard disks currently located on 24 NGAS systems (Next Generation Archiving System) [5]. The ESO archive was facing in 2005 the challenge of data volume growth to over 1 PB over the next 7 years. To meet this challenge, a Petabyte Class Archive at ESO has been designed, including the Primary Archive, an online copy with fast access of the entire dataset; the Secondary Archive, a full second copy of the Primary Archive at a physical different location; a Compute Stack with CPU, memory and bandwidth on demand including a high performance storage, the accessibility of the scientific data through internal and external network and the media production for data distribution. In addition, a complex Data Centre environment had to be designed and constructed to host and to support the various components of the Petabyte Archive in a reliable and flexible manner in respect to safety and availability.

*E-mail: dieter.suchar@eso.org; phone +49.89.32006.249

2. HISTORY

In 1998, the ESO Scientific Archive consisted primarily of 2 CD jukeboxes hosting 1000 CDs with the capacity of 600 MB each. Approximately half a Terabyte of scientific data was online on a footprint of 3 standard 19" computer racks, including the server managing the CD Archive. The contents of these CD jukeboxes were migrated to a DVD jukebox providing 1087 slots for 3.95 GB DVDs and a total capacity of almost 4.2 TB. The online DVD Archive has been gradually extended in the following years with additional 4 DVD jukeboxes of a newer type. At this stage, the ESO Scientific Archive provided a total online capacity of 16.5 TB and had a footprint of approximately 5-7 standard 19" computer racks. A single Sun Microsystems computer, based on SPARC CPUs and a Solaris 6 operating system, operated the jukeboxes acting as the retrieval part of the ESO Request Handler.

Fig. 1. This figure shows the scientific online archive in the Archive Operations Room in 2002. The DVD jukebox with its 1087 slots and 6 DVD readers on the left, another type of DVD jukeboxes with 670 slots and 4 DVD readers in the middle as well as the computer system with 4 external disks hosting the DSS, GSC and USNOA catalogs. All components were supplied through a single 3 KW Uninterruptible Power Supply on the bottom of the rack.

In 2000, an increase of the archive volume to approximately 50 TB has been expected during 2001 to 2004. A DVD Archive of this size would have required 16 jukeboxes and would have exceeded the available computer room space in the Archive Operations Room. In addition, the effort and number of personnel to produce (write and verify) the complete data inflow on compatible DVDs for jukeboxes would have gone beyond the scope of managing an archive.

To cope with the data inflow in the order of Terabytes per month in mean and some hundreds of Gigabytes during one observation night in peaks [2], it has been considered to design a new online archive on inexpensive spinning hard disks. On this background the Next Generation Archive System (NGAS [2]) has been developed and started operations in July 2001.

The first NGAS cluster consisted of 9 systems including one NGAS master unit. Each system was equipped with eight 80 GB disks and the entire NGAS system of the first generation had an overall capacity of 5 TB per rack. These NGAS systems have been in sleep mode if no request has been pending to reduce the power consumption and cooling. They

have been powered on through the master sending a wake on LAN package to a specific node. As soon as the NGAS node has received the information to become online, it has started the operating system, the NGAS server and has delivered the requested data in approximately one minute.

In the second configuration level of the first NGAS cluster, the configuration of the NGAS computers have been untouched, but the data disks have been gradually upgraded with 200 GB disks, which increased the capacity to 13 TB per rack. Most of these NGAS systems were in operations for 5 to 7 years before being retired in 2007.

Fig. 2. This figure shows the first NGAS system in the Archive Operations Room in 2002. These servers were equipped with 8 times 80 GB disks each and had a total capacity of approximately 5 TB per rack. In addition, a sequential power on procedure has been put into place for maintenance reasons and to avoid overload due to starting current.

!" #$%& '(%$)* +&,,)& -*./0$)

The basic idea for the design of a Petabyte Archive at ESO was the availability, the flexibility, the upgradeability as well as the accessibility with the overall goal of storing and protecting the scientific data. Therefore, the Secondary Archive has been installed at a different physical location and we have started to design a complex Data Centre environment to host the Primary Archive, as well as the computational and the database environments.

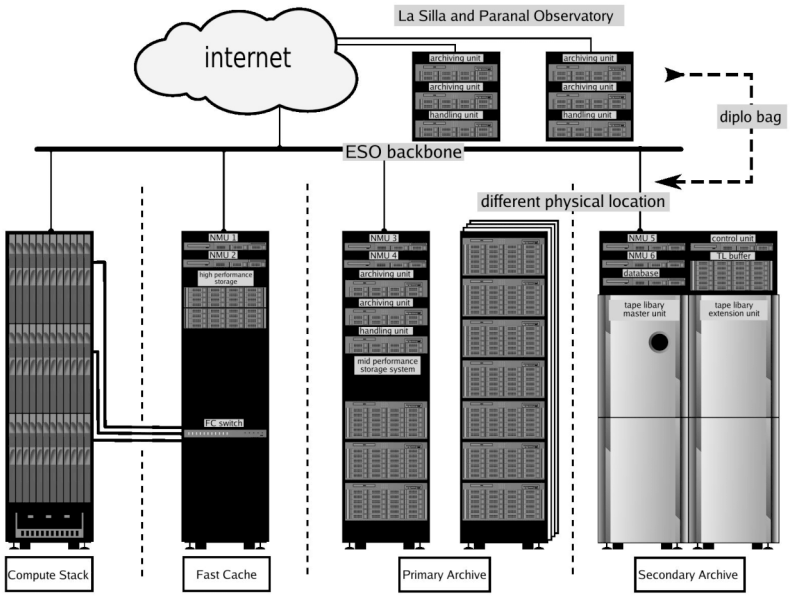

Fig. 3. This figure illustrates the operational IT infrastructure for the ESO Petabyte Class Archive. The Primary Archive consists of 24 NGAS units installed with Scientific Linux operating system and provide currently 150 TB. An entire copy of the data on the Primary Archive is transferred to the Secondary Archive. The Secondary Archive consists of a data buffer for ingestion and retrieval, a database system, a tape library based on fibre channel SAIT drives and a computer system installed with commercial software managing the entire Secondary Archive. In addition, a Blade Centre with over 100 CPUs and a high performance fibre channel storage has been installed for processing large amounts of calibration and scientific data.

3.1 The ESO Data Centre

The ESO Data Centre had to be constructed at a different physical location to separate the Primary and Secondary Archive. It has been built on the area of nine standard offices. Furthermore, it had to be modular to allow us to dismantle and move the Data Centre back to ESO premises and reconstruct the offices in the future and if needed to extend the current Data Center in size and functionality.

The ESO Data Centre is not only a state-of-the-art computer room; it is a Data Centre with leading-edge technology. Special care has been taken during the planning phase on how computer systems can be integrated into this safe, redundant and stable environment. The Data Centre can also be monitored and within certain limits managed and controlled remotely. In addition, this Data Centre has a very flexible construction matrix which allows it to easily cope with future demands.

Special attention has been devoted to the design of redundant and high capacity targeted cooling. The Data Centre has been split into two cooling environments. The part of the Data Centre, which is equipped with standard racks, is ventilated through redundant air-conditionings through the raised floor, which provides a maximum of 3-5 KW cooling per rack for non-operationally critical development, integration and test systems. The second part is exclusively equipped with redundant local Liquid Cooling Packages located next to each rack. Each LCP can in average provide up to 5 KW cooling depending on the difference between inflow and return temperature of the circulation liquid. A single rack can therefore be equipped with servers with a heat production of up to 30 KW maximum or up to 25 KW in a fully redundant setup. Also the heat exchangers including an efficient winter cooling, the pumps and the cold-water buffers have been constructed to provide full redundancy.

Particle detectors have been installed in all separated areas including the actively cooled racks, the raised floor and all standard racks in the Data Centre. Particle detectors are important components in a safe Data Centre environment and are used to alert the Data Centre personnel before a smoke or combustion event occurs. In addition, smoke detectors have

been installed throughout the Data Centre to initiate a safety procedure closing the fresh delivery and exhaust air ducts, opening the pressure sliders to avoid excessive pressure during flooding the Data Centre with fire extinguishing gas.

During the construction phase a structural engineer had to be consulted to re-compute the maximum load of the Data Centre and its IT infrastructure for the building. The IT equipment is limited to 25 tons maximum weight for the entire Data Centre. In the case of exceeding this limit, additional concrete pillars are needed to reinforce the building stability.

The measured electrical power consumption during one month is at the magnitude of 60,000 KWh, which represents an average power consumption of 400W to 500W per computer. The entire Data Centre is supplied with electrical power through a redundant Uninterruptible Power Supply (UPS) including several racks with batteries and three ACDC converters supplying 40 KVA each with currently an autonomy time of approximately 60 minutes. The ESO Data Centre is for almost one year operational and the UPS has covered several short glitches as well as a blackout.

Further equipment had to be installed in the Data Centre. Typical examples are a water detection system in consequence of water tubes and Liquid Cooling Packages in the Data Centre racks and an oxygen warning system in the fire extinguishing room. In addition, an access system and a physical intrusion system have been installed and are connected to the site security in case of emergency.

Another important requirement was a centralized system, controlling and managing the essential functions of the Data Centre. The Uninterruptible Power Supply and the Liquid Cooling Packages are managed through web-based interfaces and are controlled with a centralized Data Centre management system. This system collects all environmental information and problems with malfunctioning devices and contacts the responsible Data Centre personnel.

3.2 The Primary Archive

The Primary Archive consists of twenty-four NGAS systems providing a total storage up to 150 TB. Each rack is populated with eight NGAS systems, where each of them is equipped with twenty-four 400 GB disks, providing RAID 5 storage of 50 TB per rack. To date, the ESO Archive holds approximately 75 TB of scientific astronomical data [3]. These systems of the second generation produce a cumulative heat of 18 KW, which is chilled within fully redundant Liquid Cooling Packages across the three racks. The overall weight of these racks including the archiving systems amounts to almost three tons of hardware.

Fig. 4. This figure shows the Primary Archive in the ESO Data Centre, which consists of twenty-four NGAS systems populated with 576 400 GB disks. The second generation of NGAS systems provides 50 TB per rack distributed on 8 NGAS systems with four file systems on RAID 5.

The newest generation of NGAS systems, are operated with SATA-II Terabyte hard drives in JBOD mode. This generation of NGAS systems provides 200 TB per rack and reduces the purchase price, the cooling, the power consumption and the footprint to a quarter. These servers are equipped with two quad core CPUs because greater processing power is required perform standard archive data tasks whilst simultaneously executing the data consistency checks for the 20 TB volume on each server. This new generation of archiving systems will ensure a smooth transition to a 1 Petabyte Primary Archive at ESO which can be easily maintained. Based on current data growth the archive systems that will be necessary to maintain a 1 Petabyte Archive will not require more than 6 actively cooled racks.

3.3 The Secondary Archive

The Secondary Archive consists primarily of a tape library, a database and an application server as well as a buffer. To date, the Secondary Archive can store up to 150 TB and can be easily extended to over 1 Petabyte. In comparison with the Primary Archive: it is slower in performance, but it has a smaller footprint, less power consumption and lower heat dissipation. It keeps an identical copy of the entire Primary Archive data and is used for data recovery.

Fig. 5. The tape library of the Secondary Archive

3.4 The Computational Environment

Many internal customers at ESO depend on high performance operational computer services provided by the Operational Technical Support department. In particular, the Quality Control group is a significant user of these services. They process large amounts of calibration and scientific data from the Primary Archive in order to measure and monitor the state and the quality of the instruments, and to produce final science data products for the astronomical community [4].

Most of the operational high performance computer services are located within the Blade Rack in the ESO Data Centre. The servers within the Blade Rack are AMD based and use the Scientific Linux Operating System. Several blade servers within the Blade Rack are used as individual servers for the scientific computational environment and others are used to form a GFS shared filesystem cluster. The GFS shared filesystem allows all blade servers within the cluster to have simultaneous disk I/O access the fibre channel storage. The use of the GFS shared filesystem cluster will provide more efficient processing of data with added server redundancy and the fibre channel storage will improve disk I/O throughput.

At the hub of the archive are multiple Relational DataBase Management Systems (RDBMS), both Online Transactional Processing (OLTP) and Online Analytical Processing (OLAP). These systems: receive the proposals, aid the peer review process, transfer the observation blocks to the remote observatory sites, receive the resulting observational metadata and close the loop by tracking the location of the image and any subsequent copies.

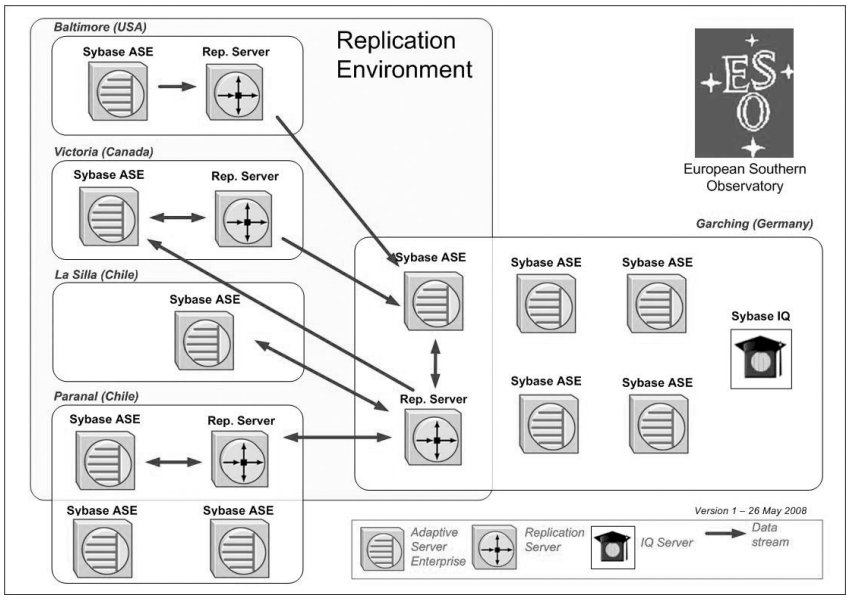

Fig. 6. The ESO replication environment

ESO has with time, increased its reliance on database technologies, it therefore became apparent that a standby system was needed to keep services running. The cold standby solution that is currently located at a different physical location, is based on a series of scripts which: take a snapshot of the production server and another of the cold standby server, look what files have not been applied on the cold standby server and then copy the files. When the files have been copied, the files are loaded but kept offline. The script is dynamic which means it will copy any new database from the production server to the cold standby server if that one exists on the cold standby too without additional coding. The downside to a cold-standby is the possibility of a small data loss should a switch have to be made, however every effort is made to limit this with transaction log dumps on a regular basis. In addition, ESO will be testing technologies that will provide a greater level of database redundancy.

To support activities in the computational environments such as: Relational Database Management Systems, Blade Centres, Web technologies to provide the gateway for users to access services and content management systems; a bespoke problem reporting system was developed to record, control and manage operational tasks for a Petabyte environment.

4. CONCLUSION

Although we are currently not yet at the 1 Petabyte mark, we are more than capable of extending to 1 Petabyte and beyond. We will be able to achieve the future challenges, because we have adopted a modular approach to the design of the Data Centre, the Primary and Secondary Archive, as well as the compute and database environments. We have created the foundations to use modern archive technologies, clustered database solutions and virtualized systems with centralized high-performance storage, as well as memory and CPU on demand.

REFERENCES

[1] Suchar, D., Technical Specification for the provision of an Archiving System, GEN-SPE-ESO-50000-3708, Issue 1, 11.07.2005
[2] DMD NGAST Initiative, ESO, http://archive.eso.org/NGAST/NGAST_index.html
[3] Fourniol, N., ESO (private communication)
[4] Hanuschik, R., ESO (private communication)
[5] Wicenec, A., Knudstrup, J., The Messenger, 129, 27 (2007)