Atacama
Large
Millimeter
Array

ALMA Archive Operations Plan

ALMA-70.50.00.00-006-A-PLA

Version: A

Status: Draft

2007-12-19

Prepared By:	**Organization**	**Date**
Gianni Raffi	European Southern Observatory	2007-12-19
IPT Leader Approvals:	**Organization**	**Date**
G. Raffi B. Glendenning	European Southern Observatory National Radio Astronomy Observatory	
System Engineering Approvals:	**Organization**	**Date**
Configuration Control Board Approval:	**Organization**	**Date**
C. Haupt	ALMA Configuration Control Board Secretary,	
JAO Director Release Authorization:	**Organization**	**Date**
T. Beasley	Joint ALMA Office Project Manager	

Change Record

Version	Date	Affected Section(s)	Change Request #	Reason/Initiation/Remarks
A	2007-12-19	All		First Draft for EDM

Table of Content

1 Description

The ALMA Archive Subsystem is a central and critical component of the ALMA computing infrastructure (see [Schwarz, 2003]) and provides generic persistence and the main data flow mechanisms for data delivery from the OSF to the SCO and to the ARCs. In addition to these data flow support functions the archive is also responsible for the long-term maintainability of the astronomical heritage of the ALMA data even beyond the lifetime of the ALMA observatory. Being in such a central role, the archive subsystem is one of the most critical subsystems along with the Control and the Correlator subsystems, but different from those two the archive has to deal with all meta-data, logging and of course the bulk data produced by the ALMA observatory. The archive will be a distributed system, which needs to be operated in a concerted, efficient and secure way at the different sites in order to be able to deal with both the data rate, the total data volume and the internal and external data requests of ALMA and the astronomical community world-wide. It should also be noted here, that the archive is absolutely essential for ALMA operations: Without the archive working properly, the computing system will not even start up, and if the archive fails or the connection to the archive is lost for some reason, ALMA will have to stop observing almost immediately. ALMA will be a unique facility covering a unique wavelength and parameter space and thus the scientific value of the data will be unique as well. The value of the data will be multiplexed by the publicly accessible science archive after the proprietary period. Also the data rate and volume of the observatory will rapidly grow to values, which are amongst the highest in astronomy. The data flow part of the archive (ALMA Frontend Archive, AFA) mostly has to deal with the data rate of 200 TB/year, while the science archive part (ALMA Science Archive, ASA) mostly has to deal with the cumulative total data volume, which will hit 1 PB after three years of full operation (including a backup copy). Given these numbers and boundary conditions the maintenance and operation of the archive is one of the most critical and challenging tasks of the ALMA observatory.

1.1 Purpose

This document is a description of the ALMA Archive operational concepts based on the planned deployment of the archive across continents and sites (see below). The archive operational concepts are dependent on a number of external and internal factors such as the role and scope of the various sites, but also the actual archive implementation and deployment. Thus this document describes boundary conditions and assumptions, the deployment of the archive and some technical details, which are relevant to understand the archive operations plan.

1.2 Scope

The document covers the commissioning, early science and the operational phase of the ALMA archive and the aspects related to data migration and long-term data

storage. The detailed setup of both the database installations and the NGAS clusters for the persistent bulk data storage is covered in separate documents [Marx, 2007], [Knudstrup, 2007]. The detailed operational procedures will be attached to those document once they are worked in collaboration with the operations staff. The current document covers the higher level operational aspects of the ALMA archive network.

2 Related Documents and Drawings

2.1 References

[Habermann, 1996] "TCP in high speed networks" http://www.tkn.tu-berlin.de/curricula/ss96/bla/tcp_hs.html

[Knudstrup 2004] "DFS Software", "NG/AMS Next Generation Archive Management System", "User's Manual", VLT-MAN-ESO-19400-2739, Issue 3, 2004-08-06.

[Knudstrup 2003] "DFS Software", "NGAS Operations & Troubleshooting Guide", VLT-MAN-ESO-19400-3103, Issue 1, 2003-07-17.

[Knudstrup, 2007] "ALMA Archive BulkStore Architecture", in preparation

[Lucas, 2004] "Estimation of ALMA Data Rate"

[Marx, 2007] "ALMA Archive Database System Architecture", in preparation

[Schmid, 1998] "QoS based Real Time Audio Streaming on IPv6 Networks" http://www.comp.lancs.ac.uk/computing/users/sschmid/Spie/paper.html

[Schwarz, 2003] "ALMA Software Architecture", Version 1.01, 2003-02-24

[Scott, 2002] "Data Rates for the ALMA Archive and Control System" http://www.alma.nrao.edu/development/computing/docs/joint/notes/DataRates2.pdf

[SCS, 2004] "Comparing Storage Alternatives for Digital Asset Management" http://www.newspapersystems.com/pres/storagealts.html

[Smeback, 2007] "ALMA Operations Plan", Version D, ALMA-00.00.00.00-002-D-PLA.A

[Wicenec, 2005a] "ALMA Archive Subsystem Design"

[Wicenec, 2005b] "ALMA Archive Subsystem Development Plan"

[Wicenec, 2005c] "ALMA Archive Hardware", http://almasw.hq.eso.org/almasw/bin/view/Archive/ArchiveHardware

2.2 Abbreviations and Acronyms

AOS Array Observing Site (Chajnantor)

ARC	ALMA Regional Center(s), the main ARCs will be located in NRAO Charlottesville/USA and at the ESO headquarters in Garching/Germany.
COTS	Common Off The Shelf, this is a term mainly used in clustering to describe the usage of cheap 'consumer-type' hardware.
DFS	ESO Data Flow System.
IPv6	Internet protocol version 6
NGAS	Next Generation Archive System, the archive file handling system, including the software and its integration into a hardware and operational concept.
NG/AMS	NGAS Archive Management System, this is the NGAS software.
OSF	Operations Support Facility (San Pedro)
QoS	Quality of service, functionality provided by IPv6 to ensure a certain quality (e.g. bandwidth) for a certain network based application (e.g. media streaming).
SCO	Science Operations Center (Santiago)

2.3 Related Interface Control Drawings

Xxx

3 ALMA Archive Deployment – The ALMA Archive Network

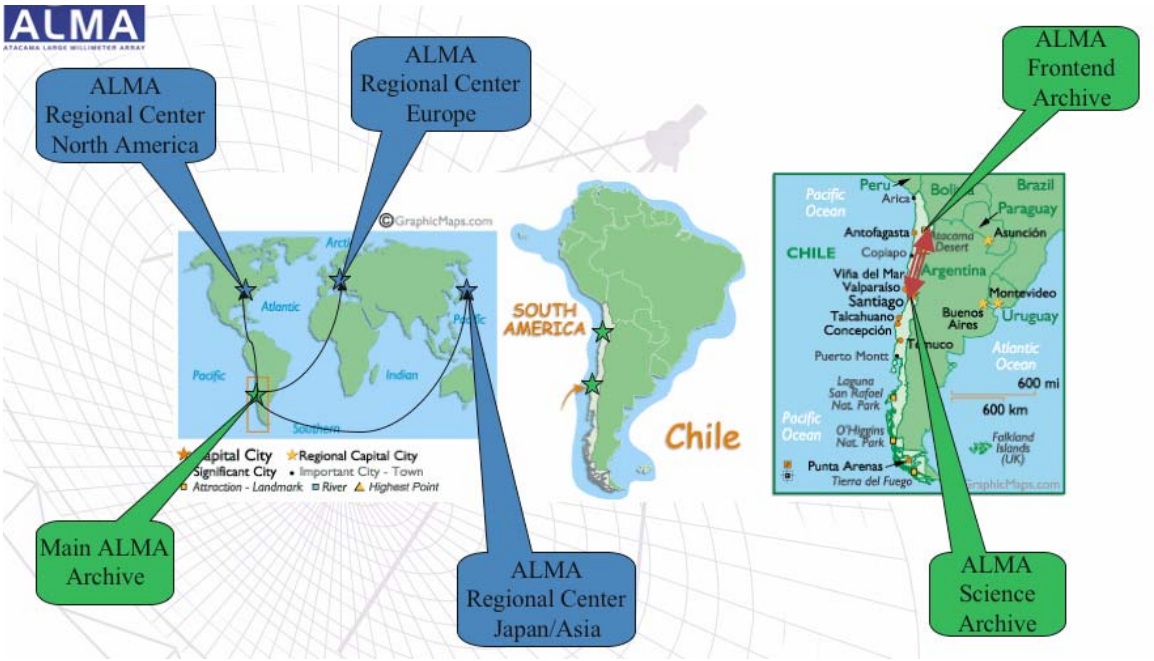

Figure 1: This figure shows the world-wide distribution of the ALMA archive network in two locations in Chile and in one location in North America, Europe and Japan, respectively.

In its final stage the ALMA Archive will be a fully distributed system, which has operational parts at all ALMA sites (except the AOS) and at the ARCs. It has a direct interface to the main data providers, Correlator(s), Control and Pipeline and provides query and retrieval interfaces for engineering and for scientific software both internal and worldwide. This wide range of functionalities will be deployed at the places depicted in Figure 1 on a range of computers in order to provide the necessary availability and performance. Figure 2 shows a conceptual view of the archive interactions both internal between the ALMA Science Archive (ASA) and the ALMA Front-end Archive (AFA) and also external to subsystems and human users. The ALMA archive design reflects the three main categories of data produced by the observatory, bulk data (correlators), XML data (project and science meta data) and monitor and log data (in general timestamped values) and thus the main components of the archive are:

1. XML store: storage of project and observation related meta data in XML database.
2. Bulk store with NGAS backend: storage of big binary data on file system.
3. Monitor store: storage of sensor data and logging data in relational database.
4. Science archive: storage and maintenance of scientific meta data in database.

5. Database: persistent database service for the whole archive.

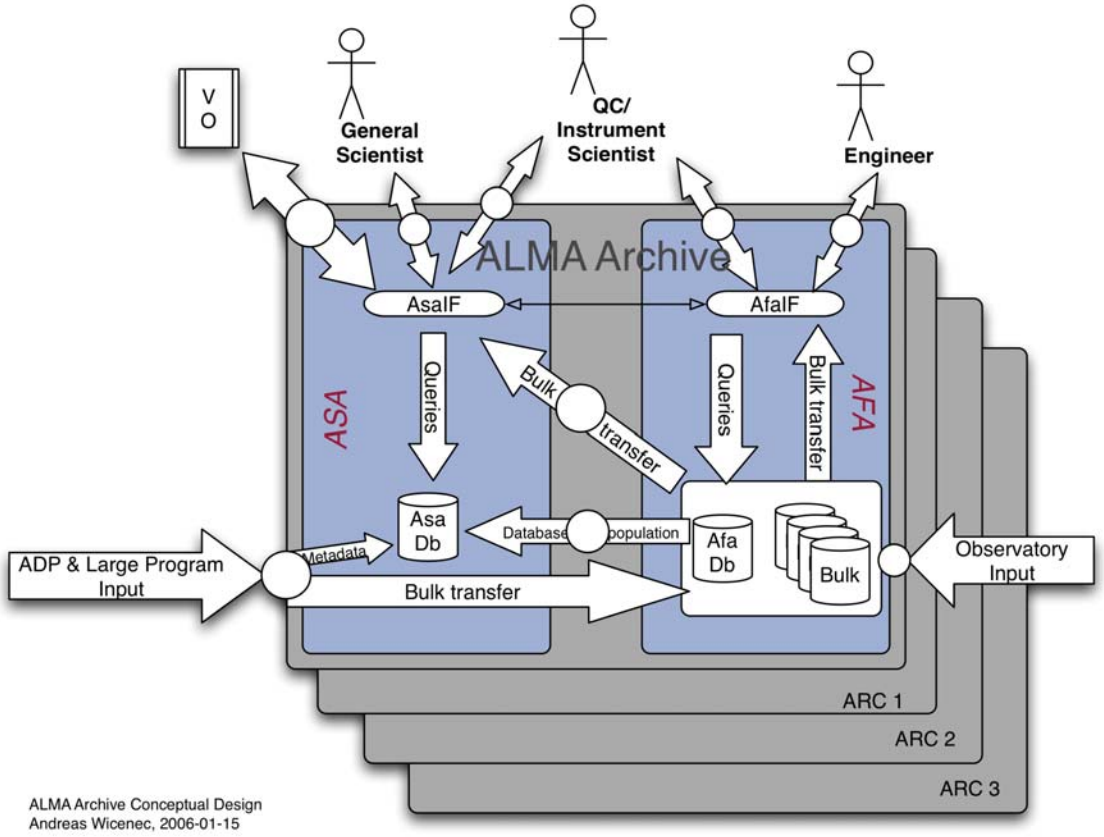

Figure 2: The figure shows the conceptual design of the ALMA archive with the major interactions outlined. The two big blocks are the ALMA Science Archive (ASA) and the ALMA Frontend Archive (AFA) respectively. The stacked boxes depict that this setup will be replicated to each of the ARCs as well. The AFA is the part of the archive providing the core persistence functionality for the ALMA data and the interfaces for the other subsystems and engineering and lower level scientific interfaces for internal human users. The ASA provides the external interfaces to scientists and VO systems. It also implements the scientific view on the ALMA data.

Where the database is the core component and is used by all the other parts and is thus the most critical part of the archive subsystem. The ASA is a separate part although it mainly shows data collected in the AFA, the main difference is the scientific orientation of the users and the requirement to support external access through the public internet. The strict separation allows us to protect the AFA and ALMA operations and also to change the quite different meta data items independently. The ALMA bulk data is not

stored in the database directly, but on NGAS, which is a file handling system providing scalable, secure and distributed handling of files of any type [Knudstrup, 2004]. Since the archive will hold data in the petabyte range within a few of years of full ALMA operations, the file handling plays a vital role in the archive operations and the NGAS operational concepts drive the outline of the main tasks to be carried out by the archive operations staff. The weight of the major parts of the archive varies between the sites and thus the tasks will also be different.

4 Archive Baselines

The operation of a globally distributed archive, which will grow to the petabyte scale within a few years of full ALMA operations, is a real challenge. This plan is based on a number of assumptions about the operational environment. The role and scope of the various sites is one key element and needs to be clarified before a fully detailed operations plan can be issued. The long commissioning and early science period poses additional operational requirements onto the archive.

4.1 Boundary Conditions and Assumptions

For this document the following assumptions are applied:

- The ramp-up of the ALMA array will be done linearly at a rate of one antenna per month.
- Early science operations starts 1.5 years after the initial delivery of the first antenna.
- We need to support the vendor tests with a minimal archive installation.
- Commissioning data will have to be archived, but the primary access point to this data will always be at the OSF.

1.1.1 Data Rate and Data Volume

- The total average data rate is 6.6 MB/s, the maximum is 66 MB/s. The total data volume for full ALMA operations will grow at a rate of 200 TB/year.
- The ACA adds not more than the nominal 7% to the total ALMA data rate (included in the numbers above).
- The monitoring and logging data does not add more than 10% to the above numbers, i.e. 0.6 MB/s with a total of 20 TB/year.
- The bulk data rate (B) scales like $N^2/2$ (including autocorrelations), where N is the number of baselines. Since the maximum data rate of ALMA is a definition of the upper limit rather than the result of a calculation this maximum will be reached earlier than with 64 antennas

with $N = \frac{N_{ANT^2} - N_{ANT}}{2}$ (N_{ANT} is the number of antennas)

we get

$$B = \frac{N_{ANT}^4 - 2N_{ANT}^3 - N_{ANT}^2}{8}$$

for $N_{ANT} = 7$, B = 220.5 and we assume that the maximum data rate for 7 antennas is 100 kB/s. The upper limit of 66 MB/s is thus 600 times higher, i.e. it will be reached for B = 220.5*600 = 132300. Resolving this for N_{ANT} leads to a number of 33 antennas. Since the assumption is that antennas will be delivered at a rate of 1/month, the maximum data rate could be[1] reached 33 month after the delivery of the first antenna.

- Depending on the scheduling blocks, the maximum data rate could be delivered for several hours, as a consequence the archive design does not treat it as a peak but rather as a plateau of arbitrary length and the plan is that the archive subsystem will be able to deal with 66 MB/s input in real-time[2] without additional buffering.

4.1.1 Number of Transactions

Another critical point is the required number of transaction per second (transaction rate). This number is critical for the database system design and the archive deployment. For the database it is the sum of all requested transactions, for NGAS it just covers the BulkStore part. The most demanding requirements here are probably coming from the monitor and logging system, where rates of up to more than 500/s/Antenna can be reached. With the currently estimated numbers we can expect a total rate of about 90,000 monitor points per second. Since such a transaction number would certainly dominate the whole data flow we (control and archive subsystems) have designed a system where not every single monitor point is one transaction, but where the monitor points are blocked in configurable time intervals.

4.1.2 Data Delivery and Interfaces

Within the distributed network of archive nodes in ALMA all data has to be replicated and transferred. The archive implementation supports both network based distribution but also media based distribution. The current baseline is to use the network for the database replication between all sites. Bulk data will be transferred by network between the OSF

---

[1] The calculation just shows that the limit could be reached; it is still dependent on the kind of projects executed whether it is actually reached.

[2] This is to be confirmed, but the raw performance numbers of NGAS show, that we could handle about 45 MB/s with the current machines using load-balancing between four nodes.

and the SCO and by media between the SCO and the ARCs. This approach is compliant with the current planning of the network availability between the different sites for the first few years. In order to be able to transfer all bulk data through the network we need at least 12 MB/s (100 Mbit/s) sustained for nominal operations and for the archive only and even then it might be necessary to catch up with sending media for longer periods of high data rates. At a first glance that does not seem to be too demanding, even for inter-continental connections, however in practice it is quite tricky to setup and tune such a long distance connection to really deliver 100 Mbit/s sustained on an open shared network. We have performed a test on a connection between Victoria (Canada) and the San Diego supercomputing center with a theoretical bandwidth of 1 Gbit/s and found a data rate of just over 10 Mbit/s. Leasing dedicated connections between the SCO and all the ARCs delivering 100 Mbit/s with current prizes is most probably out of reach (TBC), but could be feasible within 5 years or so. The archive implementation in principle allows mixing of both network and media transfers, however the management of a mixed approach naturally is quite a bit more complicated and needs more attention by the operators. Thus the recommendation is to use one mechanism only for a given connection and time interval. In general network data replication is easier to maintain than media transfer and thus certainly the preferred mechanism for the future. For the transfer between the OSF and the SCO we should certainly try to setup and tune a network connection from the beginning. Since the archive implementation can already support both network and media transfer it is possible to test both and fallback to media if the network is found to be too slow.

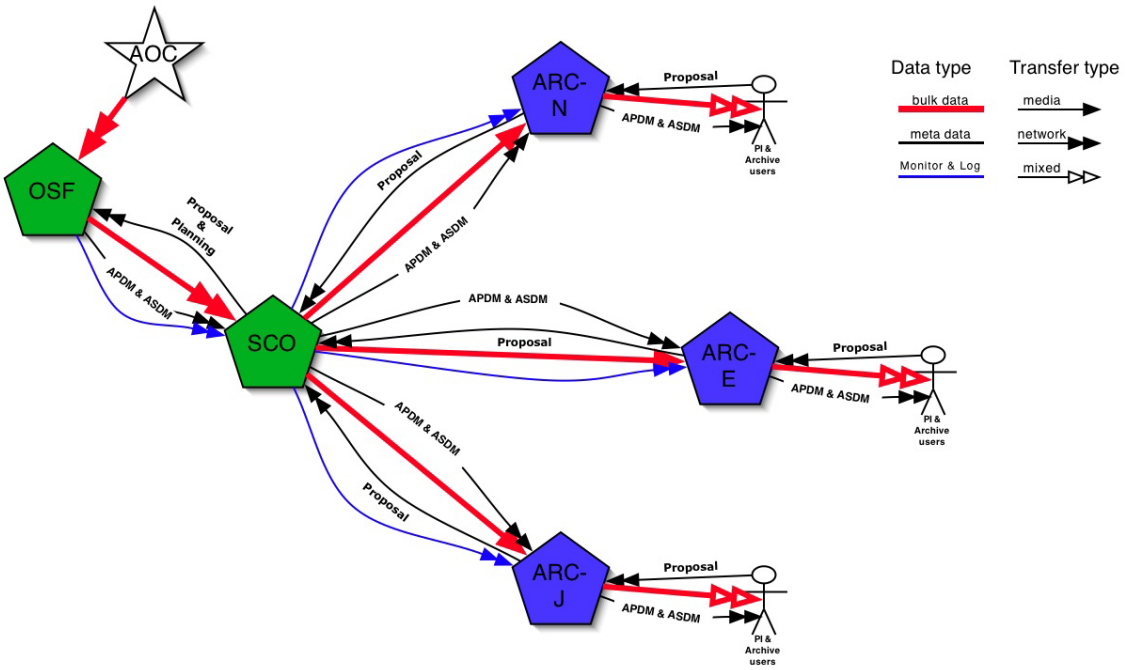

Figure 3: Schematic view of the data transfer interfaces of the ALMA archive. The figure shows all interfaces from the PI all the way to the OSF and back to the PI and the archive users. The different transfer methods and data types are depicted with different arrow ends and arrow colors,

Data delivery from the AOC to the OSF will be done through a dedicated fiber network link. This link has to host a number of services where the most demanding one certainly is the interface between the correlators and the archive. It is assumed that this link is as reliable as necessary and that the bandwidth is always high enough to ensure that all data can be delivered. By design of the computing data flow, the archive has been kept out of the quasi real-time loop required for the telescope calibration and quick-look. Thus at this stage the currently observed bulk data sets will only be archived, but not requested within a very short time interval.

Data delivery from the OSF to the SCO is assumed to be done using standard NGAS media for the bulk data, network for certain data sets (e.g. time critical science data) and for the project meta-data in particular. If enough network bandwidth is available it is possible to switch at least the average data rate to full network delivery. This can be done through configuration of the NGAS servers on both sides.

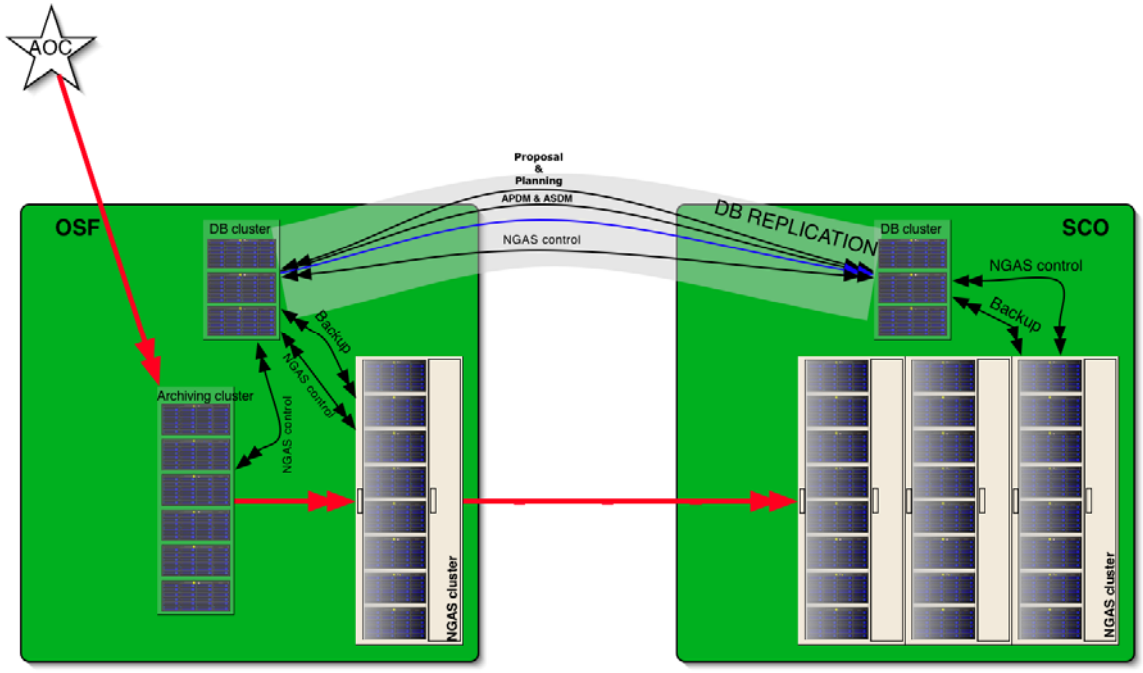

Figure 4: Schematic overview of the data interfaces between the OSF and the SCO. Arrow colors and heads are the same as in Figure 3. The wide transparent band at the top represents the database replication of various data types between the OSF and the SCO DB clusters. Bulk data is transferred from the AOC to a dedicated archiving cluster and from there to a NGAS cluster, where up to 6 month of ALMA data can be kept. As soon as the SCO is available the bulk data will also be transferred to the SCO into the main NGAS cluster. During early operations disks will be send from the OSF to the SCO, but during nominal operations the data will be transported through the network. Database backups will be archived on the NGAS cluster as well and thus the backup files

4.2 Archive Operations Outline

The archive operational tasks are directly related to the operations and maintenance of the main archive components and can be grouped into four main sections:

1. Database operations: Database monitoring and maintenance, replication monitoring and maintenance. Database backup and restore. Database tuning and configuration adjustment. Database user management and support.
2. BulkStore operations: NGAS media handling cloning and shipping, BulkStore monitoring and preventive maintenance activities and bulk data migration to higher capacity media.
3. Database content management: Monitor and maintain consistency of (meta-) data database content. Run, tune and maintain the data warehouse ingest procedures

for the ASA. Fix problems and perform schema migration activities. This activity also includes the scientific content management, like flagging and correcting data.

4. Archive user support and data handling: Internal and external archive user support, data request handling, packing and shipping of media, archive helpdesk.

Not all the activities are carried out at every site. Archive operations at the OSF will mostly be limited to maintenance, monitoring of the proper functioning of the AFA including the high availability database setup and NGAS media handling[3]. The replication of data and meta data between the OSF and SCO will be controlled as a shared task for both sites. At the SCO people will have to carry out database contents management tasks for the different stores. Also at the SCO the science archive is filled and maintained by a small group of scientists and content managers. All cloning and replication of data and meta data for the ARCs is initiated at and controlled from the SCO with some help from the remote sites (ARCs). media handling and they have to do the shipping of media to the ARCs[4]. The ARCs will mainly perform the handling of PI and normal archive user requests, user support both for scientific and . The following subsections describe the deployment and operational tasks of the archive in more detail per site.

4.3 AOC

There is no archive installation foreseen at the AOC, the archive is fully dependent on the fiber link between the AOC and the OSF. This link not only has to be reliable, but it also needs to be tuned to be able to deliver non-congested bandwidth for the correlator/archive link. This could be done using quality-of-service (QoS) functionality in IPv6 [Schmid, 1998]. The actual configuration and testing of this should be done as soon as possible at the OSF. The BulkReceiver module[5] of the archive uses a library provided by ACS implementing multimedia-streaming technology over standard TCP connections and thus the known limitations of TCP on fat pipes [Habermann, 1996] apply to the link between the AOC and the OSF.

4.4 OSF

Archive operations at the OSF mainly involves the maintenance of the availability and proper functioning of the archive persistence services provided for ALMA operations. In particular this includes:

---

[3] This is dependent on the link between OSF and SCO. If there is a high bandwidth network link, no media have to be shipped to SCO, but still operational staff has to remove full and mount new, empty media.

[4] Also this is dependent on the available link between the SCO and the ARCs.

[5] This is the software, which implements the interface between the correlator and the archive.

- Monitoring and troubleshooting of the BulkStore
- Monitoring and troubleshooting of the OSF Archive DB
- Monitoring and troubleshooting of the NGAS cluster
- Monitoring and troubleshooting of the MonitorStore
- Monitoring and troubleshooting of the Event and LogStore
- Media management
- Monitoring and troubleshooting of the OSF/SCO data transfer
- Local archive and DB support

There is a full archive installation (see section 5) planned for the OSF, this includes a database installation, the archive subsystem software and the NGAS cluster installation. In the archive design it is assumed that the OSF will keep a fully operational copy of all ALMA data, that might not be true any more after some years and since the ARCs are keeping a full copy as well, it is also not strictly required to have a second copy in Chile. Since the archive is absolutely mandatory for the ALMA operations the OSF system needs to be a high availability system. The anticipated level of availability needs to be defined by the ALMA management, but probably it should be in the range of 99.9%, which means a total downtime of less than 9h/year[6]. Also the NGAS installation at the OSF has the highest requirements in terms of availability and throughput. Since there is no large high-speed buffer between the correlators and NGAS, the front-end cluster of NGAS nodes needs to be able to archive about 66 MB/s sustained[7]. Since the data archiving rate is realized by simply adding additional NGAS nodes, a side effect will be that during periods of average data rate the redundancy level will be very high. The front-end NGAS cluster will consist of a total of 6 machines, where 4 would be sufficient to capture the 66 MB/s, the additional 2 machines are for fail over and to be able to take one machine off-line for media exchange or maintenance without jeopardizing the ability to capture the full ALMA data rate. In addition the OSF will keep at least 6 month of data (full operations equivalent) and we assume that the capacity required to keep this data (200 TB) will also be sufficient to cover the commissioning and early science phase. This requires another 8 NGAS nodes with 24 1TB disks.

---

[6] If we assume that one hour of ALMA time has a value of 50,000 $, this still means a loss of more than 400,000 $/year, and it might be worthwhile to think about spending some of this money to improve the reliability beyond 99.9%.

[7] Since the duration of the highest data rate could last several hours or even days, this is not regarded as a peak in this picture.

The OSF database installation consists of 3 machines, 2 are for the Oracle Real Time Application Cluster and one is a standby server. More details about the database deployment, the replication between OSF and SCO and the backup and disaster recovery can be found in [Marx, 2007].

4.4.1 Early Operations

At least until 2010, before the SCO archive is operational we need to keep two copies of the data at the OSF. Since the likelihood of catastrophic damages due to earthquakes is fairly high in the region of the OSF it is strongly recommended to have the two copies in two separate buildings and investigate the possibility to have a rack hosting 4 NGAS nodes in Santiago as early as possible. This would also allow to ramp up and test the data replication procedures and train people without having the pressure of a very high data rate from the very beginning. In general one important task is the tuning and adjustment of the operational procedures to the actual needs and particularities of ALMA.

1.1.2 Staffing

Archive operations at the OSF will require some kind of 'negative ramp-up'. Mostly because there is no other site active during the first two years. Other reasons are the need to settle down and tune the operational procedures and the need to train the staff. During nominal operations it is expected that we would need 1 person full time 24/7 coverage for archive operations. For the rest of this document such a 'person' will be referred to as 1 TTE (Total Time Equivalent), because the exact number of FTEs required to cover 24/7 is dependent on the exact implementation of the turno system.

For NGAS operations at the OSF it is estimated that we would need one FTE for the first two years and 0.5 TTE during nominal operations, provided that the system is running stable after the first two years.

As a summary from the database document ([Marx, 2007]) it can be said that the OSF database operations and high availability requirements requires a person with good database skills on site 24/7. The actual required fraction of a TTE is estimated to be 0.3, provided that the system is working nominally and that the person in charge has good knowledge of the ALMA database setup and of Oracle DBA tasks in general.

For both functions (NGAS and DB) during the first two years, before the SCO operations start it is assumed that there is a full FTE available at the OSF in order to adjust and fine tune the operational procedures and settle the remaining issues with the database configuration. This could be a single person supported part-time after the first year by a second person who received some training in relevant areas (NGAS or DB, respectively). The goal is to buildup the turno teams before nominal operations. The actual distribution of the work to actual persons is outside the scope of this document, but here we just outline the expected workload.

It is also assumed that during nominal operations in case of bigger problems people from the SCO will support the OSF and if necessary also be transferred within 24 hours to the OSF in order to fix the issue.

The remaining 0.2 TTE should be used to carry out database content management and reporting tasks. This includes fixing problems of data and meta data due to software, hardware or human errors somewhere in the ALMA data flow. Note that this total of 1 TTE does not include system administrative tasks of the machines or the network and it is not recommended to mix these functions.

4.5 SCO

Archive operations at the SCO includes the following tasks:

- Monitoring and troubleshooting of the OSF/SCO data transfer
- Monitoring and troubleshooting of the OSF/ARC data transfers
- Media management
- Monitoring and troubleshooting of the ASA population
- Archive QC
- Local archive support

The SCO will keep the full operational reference copy of the ALMA data and it will host the science archive. It will provide and maintain the main interfaces to the ARCs and internal archive users and the pipeline. Thus there will also be a full archive installation including database, software and NGAS clusters. The availability requirements are similar as for the OSF, because most of the data processing, the contents management and the construction of the science archive will be done at the SCO. Although it is currently not foreseen that the SCO will also act as an ARC this would be easily possible since all necessary functions are available in the software. A model where all on-line interfaces are enabled and this synchronous data requests (no media handling) are enabled at the SCO as well is doable and would only require minimal additional operational attention.

It should be mentioned here that the availability of a proper computer room with sufficient power supply and cooling and safety measures against fire is required for hosting the ALMA archive. The planning of this room has to take the quite high mass of the NGAS nodes (~120 kg) into account when laying out the basement. The cooling and power requirements have to be calculated and the room equipment has to be adjusted to them. In addition the network layout has to allow for parallel gigabit connections of many machines, else cloning and data replication to the ARCs will be network limited inside the SCO already. The details of such a planning requires a separate document and does not fit into the operations plan.

4.5.1 Early Operations

Archive operations at the SCO is assumed to start 2 years after the start of operations at the OSF, together with early science operations. As mentioned above it would be very beneficial for the safety of the data during the first two years to be able to at least store disks in a controlled environment. This could be a simple shelf in a climate controlled computer room or if possible a rack where a few NGAS nodes can be installed to host the disks. This would require some small amount of manpower.

1.1.3 Staffing

The main additional tasks at the SCO are the ingestion and maintenance of the ASA and the execution of the pipeline with subsequent quality control and ingestion of the results. The scientific tasks are quite well covered by the global ALMA operations plan [Smeback, 2007]

The operation and the maintenance of the main ALMA archive database and the maintenance and tuning of the data replication between the various sites plus database user support requires permanent DBA support. The coverage of the DBA support for the SCO should be 12/7, where people are on-call during weekends and public holidays provided that the reaction time is of the order of 1 hour. Together with the turno system of the OSF DBAs this should lead to a situation where we have at least one DBA on-call for the SCO and another one working at the OSF. This is only valid if no other database applications than the archive have to be supported. Since it is very likely that there will be more applications running on Oracle, additional DBA support has to be planned accordingly.

The NGAS operations and media preparation and handling for the data replication to the ARCs requires manual work. Since this work needs to be carried out continuously, there should be a 8/5 coverage, where we expect that a single person at any given day of the week throughout the year should be sufficient. In practice this means to have two persons sharing this activity with some other tasks.

The main archive at the SCO will consist of something like 50 computers with more than 600 spinning disks plus the required network and auxiliary equipment in full operations. Such an installation requires proper system administration. In order to keep the total number of spinning disks and machines within reasonable limits all the data has to be migrated to higher capacity disks after about 2-3 years. This requires constant monitoring of computer technology and prices and careful planning.

4.6 ARCs

The main roles of the ARCs in terms of the ALMA data flow is PI and user support. The ARCs will provide the data to 'their' users and will also provide support to their local community, both for proposal generation and for data reduction. They will keep their

own full copy of the ALMA archive. They are assumed to receive their archive copies within a TBD[8] time after the observations in an incremental way. For certain datasets priority retrieval for the ARCs is foreseen and possible. For more details on the assumptions please see section 4.1.

The deliverable from the archive subsystem to the ARCs is the archive software and a deployment and configuration setup for the operational archive. The hardware and the required database licenses have to be payed by the ARCs, but the plan is to procure them all together in order to get better conditions. This will also ensure that at least for the first phase the setup and configuration is almost identical at the sites.

5 Archive Hardware Setup

The archive hardware has to be able to keep up with the average data rate, data volume and number of transactions (database and bulk store) at any given time. In addition the archive and in particular also the hardware has to be scalable, reliable and failover safe. These requirements can be met in various ways, but we have chosen a particular one, which addresses the long development, commissioning and early science phase of ALMA and provides maximum flexibility. Due to the remote location of the OSF solutions, which depend on specialists to be available within a few hours in order to fix problems with proprietary hardware would cause very high additional maintenance costs and the availability of a stock of expensive spare parts on site. Thus the chosen concept is based on the usage of COTS hardware with medium internal and high external redundancy[9]. With such a solution both the knowledge about the components as well as the costs for spare parts is usually about a factor of ten lower. In order to limit the number of problems originating in different hardware and software configuration we are planning to deploy practically the same hardware setup with small variations to all ALMA archive sites including the ARCs. Because of the global distribution of the whole ALMA archive system the chosen hardware has to be easily available at the various sites. This means that it should be possible to procure both the hardware as well as the maintenance for the hardware locally to each of the archive installations.

5.1 Archive Hardware Deployment

All components of the archive subsystem are designed and implemented in a way to provide full flexibility of the deployment. The supported configurations range from

---

[8] This time interval is dependent on technical constraints (link) and on policy constraints, where the former will mostly define the lower limit and the latter probably the upper limit.

[9] Internal redundancy covers things like multiple power supplies and fans; external redundancy are complete machines in a warm or cold-standby configuration.

shared server installations together with other subsystems[10] to fully deployed archive installations, where the database is running on three servers, the archive front-end software with the subsystem interfaces on another set of servers and NGAS in a cluster configuration behind. The database and NGAS are fully scalable, i.e. additional servers can be added even without shutting down the rest of the system. This flexibility allows us to support almost any kind of operational scenario, from stand-alone test systems to the fully established ALMA operations. It also makes it possible to deploy the archive on-demand that means it is possible to adjust the archive to the actually required data and transaction rate. The initial deployment of the archive will thus be different from the final deployment and like this it is possible to save quite some money. In order to be able to plan the archive deployment it is then essential to get information on the data rate evolution during the next years, but the initial archive delivery should be able to support at least the first three years. This initial delivery consists of one front-end machine, two database servers and four NGAS archiving machines. Fully equipped with disks this can potentially hold about 37 TB of data using currently available disks[11]. If this turns out to be insufficient it would be a normal operational procedure to add another NGAS server. The same configuration will also be deployed to the SCO, but up to 24 month later according to the operations plan [Smeback, 2007].

The ARCs will clone the complete archive system, including the database. In this way archive operations on both sides (SCO and ARCs) would not have to worry about incompatibilities between hardware or software (databases). In the likely case that we will be shipping media to the ARCs, it is a requirement for seamless operations that the NGAS machines are as similar as possible. Since we are using COTS parts in a custom configuration the procurement has to be done using very detailed hardware lists [Wicenec, 2005c] and configuration descriptions[12]. For the initial procurement the call for tender for the hardware covering the OSF, SCO and the ARCs went out end of 2007.

Computer hardware in general, but PC hardware in particular is subject to rapid development and the availability of complete servers or certain parts cannot be guaranteed. Buying more expensive hardware might be a solution, but the price difference of procurement and maintenance fees between a COTS server and a high-grade server is pretty high and just maintenance fees might be as high as a complete replacement server. Especially during the ramp-up of ALMA it is certainly advisable to

---

[10] We are also testing VmWare installations of the full archive.

[11] In 2007 it should be possible to hold about 75 TB.

[12] This is a standard procedure for ESO and the procurement of the first ALMA archive test servers in the US was straight forward: It took half an hour to find three vendors, which are offering very similar servers, another half hour to make sure that they could provide exactly the required configuration and a couple of days to get the offers.

be able to benefit from the performance and capacity evolution. Thus the recommendation here is to replace the whole archive hardware after about 3 to 4 years. Obviously this should be aligned with major data rate jumps if there are any.

5.2 OSF and SCO installations

The two installations at the OSF and the SCO are essentially two parts of an integrated distributed system. The installations for normal operations are schematically shown in Figure 3. The current installation at the ATF and the installation for vendor tests at the OSF are just single machines running the database and also NGAS. The OSF installation for early operations initially includes 5 machines, three for the database and two for NGAS. Thus all together we need to have 8 machines at the OSF for early operations. This setup should be upgraded to the full installation in two steps shown in Figure 3 by 2012. The first step shall complete the number of NGAS archiving servers to 6, the second step shall add a rack of 8 NGAS servers Since all the early operations servers at that time will be 3 years old it is recommended to replace all of them with new machines in as well in 2012. It is expected that the old machines will still work fine and thus we could use them in the to bulk archive rack (not shown in the figure) instead of buying all of those new as well.

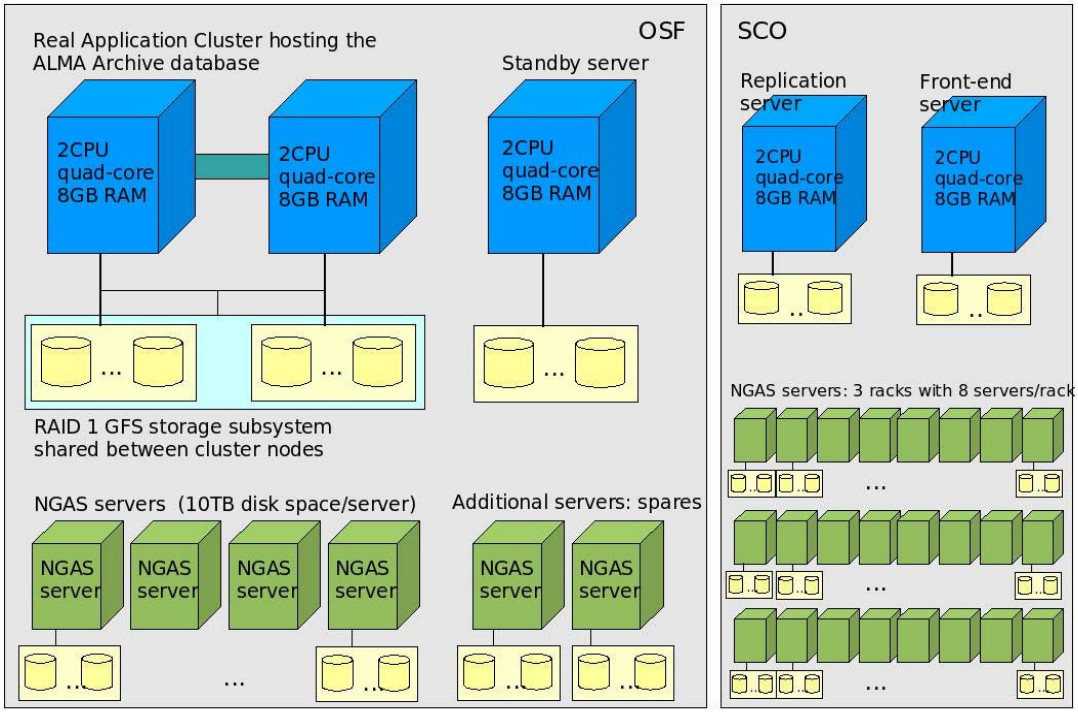

Figure 5: This figure shows the outline of the nominal installations at the OSF and the SCO. The installations at the ARCs are the same as the SCO installation. At the OSF in addition to the shown machines there will be a rack with 8 NGAS servers to hold the data of 6 month in full operations.

achieve the internal optimization there are a couple of requirements to the interface between the external and the internal components. These include:

- Dual circuit, independent power-supply chains. This will enable us to make full use of the completely redundant power supplies of the archive servers. If one power-supply chain fails, the other will keep the servers up and running, while issuing alarm signals or performing an ordered shutdown. The risk of loosing hardware due to sudden power failures or power spikes is pretty high and thus such events should be minimized.
- Dual network interfaces and switches at least for the most critical connections.
- Active hardware monitoring for the components of the most critical connections.
- Safe and reliable power-down and power-up procedures in the case of power failures and cooling failures. Reliable UPS control software, which does not shutdown machines without any reason.
- Guidelines for the power down policies (what should be available in which situation).
- Guidelines for the anticipated level of availability.

Lets assume that we need to reach a 99.9% availability of the archive subsystem as a whole (hardware and software). If the whole subsystem is running on a single server with medium redundancy, this would mean that we could afford about 4 downtimes per year with a duration of about 2.2 hours each. This could be reached with a cold, fully configured standby, 24/7 coverage with technical personnel and good replacement procedures[13]. Reasonable high availability components should have MTBF numbers about 10 times higher than the anticipated downtime, that means that every component of the system has to reach at least 10,000 hours MTBF, without taking synchronous, dependent failures into account. That does not sound very demanding and in fact all the critical components in the servers we are using have either much higher MTBF numbers, or they are redundant in the servers (power supplies, fans). The key point is the 24/7 fast reaction requirement on the technical personnel, which can be relieved by changing the configuration to a warm standby with automatic failover. Fully automatic, reliable failover on the other hand is quite tricky to implement and test. The final archive subsystem installation will work with load balancing systems with cold standby and warm standby during the periods of highest data rate. The same configuration will provide a two (database and front-end) or four-fold redundancy (NGAS) for the average data rate, respectively. Both NGAS and the database have high availability and failover

[13] For the NGAS servers the replacement of one server with a cold standby takes less than 15 minutes for trained personnel, thus we could afford up to nine downtimes within one year.

options built-in, but they need to be properly configured and maintained. It is also assumed that longer intervals of peak data rate can be planned somewhat in advance in order to be able to increase the coverage of the technical personnel.

7.1 Archive Failover, Recovery and Disaster Planning

This section gives a brief outline of the failover configurations for the databases and the NGAS clusters. The exact procedures for both failure recovery and disaster recovery have to be worked out in more detail in the future, in close collaboration with the operations personnel.

7.1.1 Failover and Recovery Procedures

The ALMA archive has been identified as a single point of failure and as a very critical system for the overall operation of the ALMA observatory. Thus the definition and implementation of proper failover and recovery procedures are essential to ensure seamless operation. The [Wicenec, 2005a] document describes the planned hardware setup of the ALMA archive installations. The proposed strategy to reduce archive failures are:

- Redundant machines with redundant parts
- Crossed watchdogs with notification
- Automatic transfer of tasks from a failed machine to a machine, which is alive.
- UPS buffering to ensure proper operation for a TBD time, but at least proper shutdown of the whole system.
- Fast replacement cycle for failed hardware.
- Fast (re-) installation, initialization and configuration of the OS, DB and the archive software.

Since all the machines used in the archive are essentially identical, they can easily be switched from one role to another and thus the whole archive can be adjusted to varying load and scale up if necessary in the future. During normal operations the redundant machines are operated in load-balancing mode. With average load the BulkStore will fully utilize just one of the four machines, thus there is a three-fold redundancy for this part. The DB will be installed on two machines in full load-balancing and failover constellation using mechanisms provided by Oracle.

Both the setup and also the automatic procedures have been chosen mainly in order to minimize the operational needs at the OSF. However it is still necessary to take actions when a machine fails and does not come back to nominal operations, or if a database node goes down. The advantage is that the required reaction time is expanded. Since

almost all ALMA data is ingested into the archive a total failure or loss of the database would be a disaster, thus both the software but also operations have to invest time and effort to make this part of the ALMA data flow as stable and reliable as possible. One very important task here is the proper maintenance of the archive hardware, including preventive maintenance.

7.2 Backup Plan

One of the top-level requirements of the ALMA archive is to keep a backup copy of all data [SSR, 2006]. This requirement has been broken up into a set of requirements in order to fit into the archive implementation and deployment.

1. The archive shall keep a backup of all bulk data.
2. The archive shall keep a backup of the XML data.
3. The archive shall keep a backup of the monitoring and logging data.

7.2.1 Bulk Data Backup

The bulk data is kept on the NGAS in files on filesystems distributed across many machines. The total amount will grow with an estimate rate of about 200 TB/year and in nominal operations there will be four copies of the data at the SCO and the three ARCs. Thus for the main part of the bulk data there will be no additional backup. However this means that we need to agree upon and setup recovery procedures for bulk data between the SCO and the ARCs in any given combination (SCO to ARC, ARC to SCO and ARC to ARC). NGAS supports such operations already, but the procedures have to be defined in detail and should also be trained.

For early operations and also for all new data, which has not yet been transferred to the ARCs or not even to the SCO it is necessary to ensure that there are multiple copies at any given point in time, preferably in different physical locations. Thus the bulk data will be transferred in a waterfall like fashion. As depicted in Figure 4 the bulk data arrives from the AOC in a dedicated NGAS front-end cluster, which is able to capture the data with the required maximum speed of 66 MB/s. This cluster of 6 machines will have a capacity of 120 TB during nominal operations and could thus store up to more than 10 days of data at the maximum speed in two copies. The data will be picked up by the main archive cluster at the OSF with about four times the average data rate and is thus 2.5 times slower than the front-end cluster. For safety reasons it is recommended to split the main archive cluster into two parts and locate them in two different buildings during early operations. The main cluster configuration can then be adjusted to produce the two additional copies in those different buildings. Once the data is safely stored in the main OSF archive cluster it is marked obsolete in the front-end cluster and will be deleted automatically by the NGAS software janitor thread. As long as the network between the OSF and the SCO is not capable of handling the average data rate, one of the copies (full

disks) can then be sent to the SCO. Once the 155 Mbit/s link between the OSF and the SCO is available the second copy at the OSF can be moved to the SCO. If the network link is not available or if the data rate is very high for an extended period, switching back to disk based data transfer is easily possible. The data at the OSF main archive cluster will also be marked obsolete once the data has safely been stored in at least the SCO and one additional ARC and will then be deleted by the janitor thread as well. Since the main archive is designed to be able to store up to 6 month of data during nominal operations, the data has to be at the ARCs before.

In this picture the front-end archive and the main OSF archive cluster are actually cache archives with a limited capacity. Only the main SCO archive and the ARCs will hold all the ALMA data. Data transfer to the ARCs will be done using disks. If one or all of the network links between the SCO and the ARCs is capable to deal with the average data rate it is possible to use the same network based bulk transfer mechanism as between the OSF and the SCO.

7.2.2 XML Data Backup

All XML data in ALMA will be stored in the Oracle database. Since the whole Oracle database will be backed up using standard Oracle backup mechanisms, there is no special backup procedure foreseen for the XML data (for more details see [Marx, 2007].

7.2.3 Monitoring and Logging Data Backup

As above this will be covered by the Oracle backup procedures described in [Marx, 2007].

7.3 Disaster Plan

In the case of a complete loss of the whole OSF archive (both copies), the amount of data lost is dependent on the delay of the synchronization between the OSF and the SCO. Due to the deployment of the archive on many machines, two buildings and the distribution of the data on many media, the reason for such a disaster must be a really destructive event, which will most probably destroy most of the computing infrastructure at the OSF[14]. For the archive the recovery is comparatively simple, but might be time-consuming, depending on the operational resources available at the SCO. Short term (~1 day) the SCO could configure two machines with a minimal part of the archive and the software installation and send them to the OSF. This would allow to resume operations for programs delivering the average data rate. The full recovery would involve the replication of all data in the SCO and should start immediately with the replication of the most recent data, which has not been delivered to any of the ARCs yet[15]. Assuming that

---

[14] The archive shares the same room with the rest of the computing hardware.

[15] These data would be available in a single copy only.

we could get an empty NGAS machine for every NGAS machine in the operational SCO archive, the replication could be done in a one-to-one configuration and would take about 3 days with currently available hardware and independent of the total amount of data in the archive[16].

The recovery plan for the complete loss of any other copies of the ALMA archive (SCO or ARCs) is similar. The sequence of recovery has to obey the following rules:

- Highest priority have those data which after the failure only have one copy left.
- Data replication must get the highest priority for the archive operations at all sites involved.

8 Archive Maintenance and Upgrades

A petabyte class archive needs proper maintenance and regular hardware upgrades in order to keep the data accessible and maintainable. Any hardware older than about 5 years tends to become in-economic for various reasons:

- Failure rate increases.
- Media aging.
- Maintenance contracts get more expensive, because hardware is not supported anymore.
- Hardware is not supported anymore by third party products.
- Capacity much lower than with modern devices/media
- Power consumption much higher than with modern devices.
- Performance much lower than with modern devices.
- Operational costs are proportional to the number of media, i.e. lowering the number of media lowers the operational costs.
- The physical volume of the whole archive gets too big.

ESO's experience shows that the turn around time of a certain media version is around 2.5 years and the turn around time of complete storage technologies is around 7 years, i.e. every 2.5 years the archive needs to change from one version of a certain media to the next (follow the road-map). Usually this involves the upgrade of reading and writing devices as well, but it is not necessary to move all the data to the new media, because the new devices are in general backwards compatible. After about 7 years the whole

---

16 This does not take the configuration and operational overhead into account.

technology becomes obsolete and some other technology takes over. The reason for such a change could also be that some technology gets so much cheaper that it simply is more economic to use that one. An example of these scenarios is the evolution of the ESO archive from WORM media to CDs and DVDs and now magnetic disks. The transfer from WORM to CDs and from DVDs to magnetic disks can be seen as technology changes due to economic reasoning. The transfer from CDs to DVDs and from one capacity version of DVDs to the next can be seen as following the road map, although the CD/DVD transition involved quite a bit more than just that.

Since the technological development (capacity) roughly follows Moore's law, which in fact is exponential, after 7 years it is even more economic to transfer all the data[17] of the archive to some modern technology.

It should be noted here that since this transfer involves only the transfer of the bulk data, it must be ensured that the references between bulk data entities and from meta-data to bulk data entities are strictly preserved during the migration process, else the data is essentially lost. In order to achieve this the archive implementation completely separates file handling and referencing from the meta-data layer. This means that all entity references are Uniform Resource Identifiers (URIs), which get translated to Uniform Resource Locators (URLs) by the bulk storage layer upon request only. For archive operations this means that by using the proper mechanisms and procedures provided by the bulk storage layer to copy, move or migrate data the integrity of the archive is preserved and thus the data stays accessible.

9 Archive Costs

The archive costs split in two main categories, the running costs to support the inflow of new data from the ALMA observatory and the maintenance costs to keep all of the data holdings alive and accessible. In this case maintenance costs also mean costs to adjust the archive to new technologies, which may involve software development and staff training. If the archive follows the technological development the average costs can be kept roughly constant as long as the data rate does not increase faster than Moore's law. The reason for this effect is that commercial hardware tends to cost always the same if a certain quality/performance/capacity level is kept. For example if you pick a price now for a magnetic disk you will get about twice the capacity in roughly two years for the same price[18], this is simply due to market laws, where you can only sell products up to a certain maximum price.

---

[17] Even tough the transfer speed does not increase with the same rate, 7 years are long enough to make such a transition possible.

[18] This is expressed by the formula:

9.1 Initial Costs

The initial archive costs are split into the pure hardware costs and the license costs for the Oracle database. The prices have to be negotiated and as of end 2007 we have issued two calls for tender, one for the hardware and one for Oracle licenses. The actual offers will give us a clear picture about the initial costs of the archive per site. Since the requirements for the ALMA archive and the initial installation are not expected to change significantly, the maximum price relation will help us lowering the hardware costs[19].

9.2 Operational Costs

The archive operational costs are covered in the [Smeback, 2007]. It should be mentioned here that every site needs a 24/7 coverage for the DB admin and archive operator roles. In addition every mirror archive site needs to be supported (media and network link handling and monitoring of the mirroring process), this requires 0.5 FTE/site at the SCO. The archive quality control, archive maintenance and request media handling requires 6 FTEs, where 2 FTEs should share their time between the SCO and the OSF. At least twice a year 2 persons should travel to the ARCs and regular ALMA archive meetings involving all people should be scheduled as well.

The hardware operational costs are hard-drives and additional NGAS nodes to capture the data from ALMA operations and to maintain the bulk data transfer to the ARCs and big data requests from power users. In addition DVD writers[20] and DVD media plus and USB/Firewire disks to support bulk data requests of average users, which are not able to download the data through the network.

In addition there are maintenance and upgrade costs for software licenses, mainly for Oracle.

9.3 Upgrade Costs

Archive upgrades of the 'technology change' type described in section 8 may require some initial investment, but on average the costs for upgrades should be covered by the annual budget, which should be kept roughly constant. If this is the case then the maximum market price relation (see footnote 17) can be used to calculate the optimum time, when to transfer to newer media/technology and if there are no other constraints, which have to be factored in, the updates could be done cost neutral. With the current

---

$price_T/price_{T0} = capacity_T/ (capacity_{T0} * 2^{(T-T0)/MT})$

with (T-T0) is given in month after the initial procurement and MT is the doubling rate also given in month. This is even true without taking inflation into account

19 One year delay in procurement saves us approximately 30%.

20 The usage of tapes should be discouraged, because tape exchange is very error prone.

price technology development it appears that following the road-map for a certain technology should be done in time frames of about 2 years, while still taking advantage of the more continuous price development[21]. Changes of storage technology should be carefully planned and the new systems have to be able to support the data migration very effectively and as automatic as possible. There should be somebody in charge of monitoring the technical development and the market situation. Yearly reports, which include sensible comparisons of the available storage technologies, should be provided. If a full technology change is recommended, resources should be made available to plan and carry out the installation and the data migration.

[21] That means that buying large amounts of media upfront for a long time should be avoided, since the prices are constantly falling for a given product.