MAGIC-TDAS 03-03

000303/RPaoletti

The Data Strategy of the MAGIC Experiment: Requirements on Data Archiving and Storage

M. Merck[a], R. Paoletti[b]

a. Institut für theoretische Physik und Astrophysik, Universität Würzburg, Germany
b. Dipartimento di Fisica, Università di Siena and I.N.F.N. Sezione di Pisa, Italy

This document summarizes the technical requirements for the data archiving and storage of the MAGIC telescope and the infrastructure for the MAGIC data strategy. The specifications for the data format and the requirements for the analysis and reconstruction software are described elsewhere [1,2,3,4].

1 Introduction

The MAGIC telescope collaboration is building a very large atmospheric imaging cherenkov telescope, with a mirror surface in excess of 230 m^2. With this detector it will be possible to detect cosmic gamma-rays at an energy threshold lower than any existing or planned terrestrial gamma-ray telescope (~ 30 GeV in Phase I and < 15 GeV in the future).

The MAGIC international collaboration is composed by institutions from Armenia, Finland, Germany, Italy, Poland, Spain, Switzerland, Ukraine and the United States of America. The telescope itself is located at the Observatorio del Roque de Los Muchachos on La Palma, Canary islands, Spain. However due to the limited personnel available on the island, the limited internet connection and the remoteness of the area, a Central Data Center (CDC or Tier 0 center) will be established. The CDC will handle the generation of data summary tapes (DSTs) and make all data of the MAGIC experiment available for the whole collaboration and specifically to smaller and extra-european partners, who cannot afford the necessary computer infrastructure. At several regional sites of major collaboration partners (Italy, Spain, Switzerland) Regional Data Centers (RDC or Tier 1/2 centers) will be built to offer the best possible access to data subsets for the physics analysis.

MAGIC will observe gamma-ray events with a camera of 577 photomultiplier pixels. Each pixel has a high-gain and low-gain channel connected to a 330 MHz 8-bit Flash ADC. Events will be recorded after a trigger signal. The expected trigger rate will be

around 1 kHz. Each event will consist approximately of 20 Kbytes (577*30 + 15% for headers). This gives a data rate of 20 Mbytes/s during full data taking.

MAGIC will mainly operate during moonless nights. The length of moonless nighttime at La Palma varies over a month from 0 to 7 hours during summer month to 0 to 11 hours during the winter period. From the 220h of astronomical nighttime during June and 320h of nighttime during December, approximately half of this time will be moonless. This means that an average of 135h of moonless night hours will accumulate each month or 1600h over one year. An additional 25% of data is expected as MAGIC will also operate under partial moon conditions (crescent far away from observational target). This adds up to 2000h per year of data. At a rate of 20 Mbytes/s this gives a total raw data amount of 144 Tbytes per year. Even taking into account a loss of data due to bad weather of 20%, this will leave us with 120 Tbytes of raw data per year. This amount of data cannot be archived given the financial constraints in the collaboration. In a first step the data must be reduced by a factor of 10-20. After this task some 8 Tbytes of data will be generated each year. The data center will be designed to hold this amount of data accessible online and store the reduced data onto tapes for longtime storage.

Before the data are transferred to the CDC they must be copied during daytime. Given the maximal length of moonless nights of 11 hours and the data rate of 20 Mbytes/s we will need to save data during daytime at a maximal tape transfer rate of 40 Mbyte/s to allow for 2 copies to be archived. One copy will stay in La Palma, whereas the second copy will be transferred to the CDC. Upon successful reading and preprocessing (suppression of noise, calibration procedure etc.) both copies of raw-data will be reused. The compressed data will be stored on tape for long-term archiving while being made available in the mid-term storage area for further analysis.

Ech month pre-processed DSTs will be transferred to the RDCs. The use of these tapes at the RDCs may vary on the analysis being performed, like studies of variability of sources, comparison of individual sources, search for pulsar frequencies, etc. For special studies, e.g. background and trigger efficiency analysis, the RDCs may have access to copies of the raw data tapes.

2 Raw Data Archiving

Raw-data archiving will be performed during day-time in La Palma. The already existing online data acquisition system (DAQ) consists of a multiprocessor INTEL-based server running the Linux operating system. The data are acquired during nighttime and stored on a RAID-0 disk array. During daytime two identical copies are written to tape. Due to redundancy and maintainability reasons the system shall consist of 2 identical subsystems. Each subsystem should be able to handle at least 66% of the maximum rate. The tape systems must be connected to a dual channel LVD SCSI card using VHDCI connectors.
After shipment of the tapes to CDC these will be sequentially read and preprocessed. For this purpose a tape library connected to a server will be used. As this library can be operated continuously the transfer rate requirements are more relaxed as for the La Palma systems. After preprocessing a dual long term archiving copy will be made and stored for future reference. The data will then be transferred to the main archive for

scientific processing. The main archive will be accessed by scientists from a cluster of Linux based workstations. The workstations are connected by 100 Mbit/s switched Ethernet connections to the university ethernet backbone. The individual workstations shall access the archived data over a mounted NFS filesystem. The Archive-Server is responsible to translate from the physical archive to a logical NFS filesystem. Results from the physics analysis will be stored in the archive or partly on local disks. Backups of the results from both, the archive or local workstations shall be possible. The backups shall be done using the same tape library as used for data import.

3 Technical Requirements

3.1 Tape Library – La Palma

A redundant system consisting of 2 identical subsystems is needed for the La Palma site. The following specifications are for the combined system

Storage Capacity:	> 1.6 Tbyte (uncompressed)
Write throughput:	> 40 Mbyte/s (uncompressed)
Host Interface:	LVD SCSI
Host Interface performance:	> 40 Mbyte/s
Mounting:	Rack mountable
Power source:	200-240 VAC @ 50Hz
Operating Temperature:	+10° to +40°C,
Humidity:	5-85% (non-condensing)
Altitude:	0m - 3000m

3.2 Tape Library – Central Data Center and Regional Data Centers

Storage Capacity:	> 1 Tbyte (uncompressed)
Read/Write:	> 20 Mbyte/s (uncompressed)
Host Interface:	LVD SCSI
Host Interface performance:	> 40 Mbyte/s
Power source:	200-240 VAC @ 50Hz
Operating Temperature:	+10° to +30°C,
Humidity:	20-80%
Altitude:	0m - 2000m

3.3 Storage Archive – Central Data Center

Storage Capacity:	> 8 Tbyte
Sustained Transfer Rate:	> 50 Mbyte/s (uncompressed)
Power source:	200-240 VAC @ 50Hz
Operating Temperature:	+10° to +30°C,
Humidity:	20-80%
Altitude:	0m - 2000m

3.4 Archive Server – Central Data Center and Regional Data Centers

Operating System:	UNIX / Linux
Filesystem access software:	NFS
Network connectivity:	1 x 1 Gbit Ethernet 1 x 100 Mbit Ethernet
Power source:	200-240 VAC @ 50Hz
Operating Temperature:	+10° to +30°C,
Humidity:	20-80%
Altitude:	0m - 2000m

3.5 Computer clusters – Central Data Center and Regional Data Centers

File server	> 1 (e.g. SuperMicro)
Disk storage	RAID-5 disk array
Disk capacity	> 2 TB, SCSI or EIDE
Number of computer nodes:	4-10
Processing power per node:	> 2 x 2.8GHz Intel Xeon or AMD Athlon
Memory per node:	> 1 Gbyte
Network connectivity:	1 x 1Gbit Ethernet (switched connection to archive server)
Operating system:	Linux
GRID software environment:	CONDOR / Globus

3.6 Software

La Palma Backup:	Linux supported tape backup (`tar`) and robot operation (`mtx`).
Central Data Center:	Library robot operations and sequential restore of data coming from La Palma.
Regional Data Centers:	Library robot operations and sequential restore of data coming from La Palma.
Archive Server:	Operation of the archive. NFS based access to the data stored on the archive. Network backup functionality.

4 Detailed Requirements

4.1 Tape Technology

The MAGIC collaboration envisions the use of LTO Ultrium technology based tapes and drives for data storage. The high tape capacity combined with the high throughput of individual tape drives and the low price per Gbyte of the media are the main reasons for this selection. A switch to LTO-2 technology, as soon as new drives will be available, will take place.

REQUIRED: ≥ 15 Mbyte/s streaming transfer rate.
< 1$/Gbyte (uncompressed) tape cost when buying 10 tapes or more.

4.2 Tape Libraries – La Palma

A system allowing to make two copies of as much as 800 GB in less then 10 hours is needed. The systems shall be connected to the DAQ system using LVD SCSI cables with a VHDCI connector on the host side. All cables necessary for the connections to the DAQ host shall be provided.
For redundancy purposes the system should be designed to consist of two identical subsystems. Each subsystem may consist of more individual components. In case of total failure of one subsystem the remaining system should be able to store 66% of the maximum amount of data (1200 Gbyte) in 12 hours. This will still allow to make 2 copies of the data for the majority of nights and a full copy for the few very long moonless nights.

The backup of data will be done using the `mtx` and `tar` commands under Linux. Operations of the tape robot using the `mtx` command shall be proven for the proposed solution. The environmental specifications for the tape libraries should allow for operations at 2500m altitude and at very low humidity levels of 5%.

The adopted solution for the data writing is to use two tape libraries in parallel with two drives and 20 slots each. A disk server with SCSI disks in RAID-0 configuration with two SCSI Ultra Wide SCSI channels to libraries.
The best candidate for the tape library is the StorageTek L40. Every library can mount up to 4 drives and a maximum of 20 tapes.

Storagetek L40 tape library

1-4 tape drives, 20 cartridge slots
L0402RH frame – 8810 $
TLTO002 tape drive – 6400$

4.3 Tape Libraries – Central Data Center and Regional Data Centers

A system allowing to load the tapes coming from the observatory at La Palma in the Data Centers is foreseen. The library should be able to read the tapes as written in La Palma. If this library is not directly connected to the main archive server a dedicated server which connects to the library has to be implemented. The preferred way of reading the data will be again the `mtx` and `tar` commands.

It is desirable to have the same type of tape libraries in the Data Centers as in La Palma. This would make the Data Centers library a hot-spare system for the La Palma operations in an extreme catastrophic event.

5 Detailed Requirements

5.1 Central Data Center

The main archive shall store up to 8 Tbyte of data in a redundant way. The data shall be accessible to the users as an exported NFS filesystem. Two different approaches are viable. A big RAID-5 disk array or a big tape library using an HSM software and an intermediate staging disk system. In case of a HSM based library system redundant tape copies are needed. Access time and read performance are not very critical as the main analysis will normally read data sequentially. However the flexibility of a RAID-5 system will be favored if it can be purchased at a competing price.

Full redundant, hot-swap, power supplies, fans and controllers for the archive should be used. A concept for a future extension of the archive to bigger capacities up to 20 Tbyte, either by adding hardware or exchanging drives or tapes by newer technologies, shall be provided. SAN connectivity may be used as on option.

5.2 Regional Data Centers

The main archive shall store up to 2 Tbyte of data in a redundant way. The data shall be accessible to the users as an exported NFS filesystem. A big RAID-5 disk array is foreseen. In case of a HSM based library system redundant tape copies are needed.

Full redundant, hot-swap, power supplies, fans and controllers for the archive should be used. A concept for a future extension of the archive to bigger shall be provided.

6 Archive Server

6.1 Central Data Center and Regional Data Centers

The server must offer the NFS protocol and run a standard UNIX/Linux operating system. All connectivity to the main archive must be available. The server, hostbus adapters and operating system must be a supported configuration by the archive manufacturer. The operating system must be installed on a hot-swappable mirrored disk. Redundant power supplies in the server are required. A journaling filesystem shall be used for the archive or staging disks of the archive to allow for a fast recovery in case of a server crash. User friendly administration tools for the archive and/or HSM system shall be available. All necessary licenses for volume managers, filesystem software and archive software including the HSM software, if applicable, must be included in the offer together with a software maintenance contract for at least 5 years.
The archive server shall be connected to the ethernet backbone of the Data Server by a 1 Gbit ethernet connection. A second 100 Mbit connection shall also be available and configured as an optional pass.

As the local Workstations will run the Linux operating system with the SuSE or RedHat distributions it is desirable to have the same operating system on the server. This is however not a mandatory requirement and more weight will be given to a configuration which is supported by the original manufacturers of the hardware

components. Official support for all components, in the installed configuration is desired.
The server should be extendable to improve performance if needed or to allow for an expansion of the archive. All upgrade options shall be described in the offer.
The server may also be used to run a DBMS system (not yet decided but most probably ORACLE) which will store all relevant information from the MAGIC operations and analysis.

A dedicated RAID-5 system of approx. 0.5 Tbyte for the home directories of the users also connected to the server may be used. Backup software for the home directories using part of the tape library shall be used in this case.

6.2 System Installation

The whole setup for the Data Center shall be installed by the host institution. The host institution will provide climatized floorspace. The tape library, the archive and the server will be installed by the host institution. It is also responsible of installing all operating software needed in order to run the archive as detailed above. Full system documentation must be provided. All costs associated with the system installation and one-day instruction of the local personnel must be provided by the host institution.

7 Regional Data Centers

Individual institutions in the MAGIC collaboration can get organized in regional data centers where Monte Carlo data are generated and/or selected streams of data are stripped, calibrated and reconstructed. The host institution is responsible for providing an adequate climatized room where the computers are located, install the needed software (e.g. simulation and/or reconstruction software). In any case, the data must be available to the whole collaboration.

8 Share of Responsibilities

The institutions in the MAGIC Collaboration agree to share the responsibilities as summarized in this table. It is left to the individual institutions the task of cost estimation and coordination with other partners for the request to their national funding agencies.

Institution	*Share*	*Responsibility*
M.P.I. München	50%	La Palma Data Center
	25%	Data Validation
University of Würzburg	100%	Central Data Center
I.F.A.E. and U.A.B. Barcelona	25%	Monte Carlo production
	25%	Data validation
I.N.F.N. Padova	30%	Monte Carlo production
I.N.F.N. Pisa	50%	La Palma Data Center

	25%	Data validation
I.N.F.N. Udine	10%	Monte Carlo production
Universidad Complutense de Madrid	10%	Monte Carlo production
ETH Zürich	25%	Monte Carlo production
	25%	Data validation

References:

1) MAGIC-TDAS 02-08, 020806/WWittek
   Proposal for the standard analysis of MAGIC data
   R.K.Bock, W.Wittek, 06 August 2002
2) MAGIC-TDAS 02-01, 020101/OBlanch
   Description of the root data format for the MAGIC telescope
   O.Blanch, 1 January 2002
3) MAGIC-TDAS 00-09, 000808/JCortina
   Description of the DAQ raw data format
   J. Cortina and J.A. Coarasa, March 23, 2000.
4) MAGIC-TDAS 00-03, 000303/H.Kornmayer
   Requirements for the MAGIC Analysis and Reconstruction Software
   J.C. González, H. Kornmayer, March 3, 2000