First thoughts on KM3NeT on-shore data storage and distribution facilities

M. Stavrianakou[a]

[a] *formerly at NESTOR Institute for Astroparticle Physics, National Observatory of Athens, 24001 Pylos, Greece*

Elsevier use only: Received date here; revised date here; accepted date here

Abstract

The KM3NeT project studies the design of an underwater neutrino telescope combined with a multidisciplinary underwater observatory in the Mediterranean. Data from the telescope will arrive on shore where they will be processed in real time at a Data Filter Farm and subsequently stored and backed up at a central computing centre located on site. From there we propose a system whereby the data are distributed to participating institutes equipped with large computing centres for further processing, duplication and distribution to smaller centres. The data taking site hosts the central data management services, including the database servers, bookkeeping systems and file catalogue services, the data access and file transfer systems, data quality monitoring systems and transaction monitoring daemons and is equipped with fast network connection to all large computing sites. Data and service challenges in the course of the preparatory phase must be anticipated in order to test the hardware and software components in terms of robustness and performance, scalability as well as modularity and replaceability, given the rapid evolution of the market both in terms of CPU performance and storage capacity. The role of the GRID would also have to be evaluated and the appropriate implementation selected on time for an eventual test in the context of a data challenge before the start of data taking.

PACS*:* 95.55Vj, 01.30Cc

Keywords:KM3NeT; GRID; Data Storage;

1. Introduction

The KM3NeT telescope [1] host site, i.e. the site closest to where the actual detector shall be installed and data taking shall take place will inevitably play a key role in the acquisition, management and distribution of the experimental data. It shall host not only the Data Acquisition, hence referred to as DAQ, but also the Data Quality Monitoring services, the Data Filter Farm and central data management services. The latter include database servers, tape vaults and robots, bookkeeping systems and file catalogue services, data access and file transfer services, data quality monitoring systems and transaction monitoring daemons. The Data Filter Farm shall run the calibration, triggering and event building and optionally part of the standard reconstruction on a data

subset. For the data distribution to the participating institutes, the site shall have to be equipped with a fast network connection to all the major computing centres. Depending on the computing agreements and the availability of large centres in the participating institutes, the above on-shore infrastructure can be minimal in terms of computing power. However, in order to guarantee data persistency and transfer, it should be adequate in terms of storage capacity and network connectivity.

Finally the site shall also be hosting the associated sciences DAQ and computing centre and must offer data processing, management and distribution services similar to those for the neutrino telescope. The site should guarantee the smooth and efficient running of the above services and assure the integrity of the data as well as their timely transfer to the collaborating institutes.

The vast experience garnered during the preparation of the Large Hadron Collider (LHC) experiments at CERN and in particular that of CERN itself as the host site with the central computing, data management and distribution centre, may serve us well, albeit bearing in mind the different target scope and scale of KM3NeT. In particular, the LHC Computing Grid preparation [2] experience will come in handy when looking for tools and services and exploring compatibility, scaling, cost and support issues.

2. Data arrival and storage

According to one scenario under consideration in the KM3NeT Conceptual Design Report, hence referred to as CDR [1], all data are transferred to shore at a rate of ~1-10 Gb/s per DAQ output node. The data are processed in real time by the Data Filter Farm for calibration, triggering and event building.

Pattern recognition algorithms based on space-time relationships acting on a snapshot of the data of the whole detector are expected to reduce the background rates by a factor of 10^4 to 10^5. The process involves calibration using local and extended clusters in the detector and is followed by Event Building: when the data pass the triggering criteria an event is built from information from all optical modules in a time window around the hits causing the trigger. Following the above processing, the output data rate should be of the order of ~100 kb/s per DAQ node.

Output data are stored on Filter Farm disks, an operation which should be sustained over a short period (order of 10s to minutes) of data taking, in addition to concurrent transfers to temporary or semi-permanent storage on volumes adequate for several weeks of data taking. This temporary, local transfer and accommodation on to Filter Farm disks would act as a safeguard against potential network bottlenecks or failures.

The above scenario raises a crucial and timely question, namely whether really “raw” data are and should be immediately discarded and thus permanently lost. A storage system whereby at least part of the data are saved for further studies of backgrounds and trigger efficiencies may have to be evaluated in depth before the start of data taking.

3. Data organization

Experimental data are typically categorized by origin. One refers to event data to describe collections naturally grouped and analysed together as determined by the level of data processing: trigger, “raw”, “Filter Farm Data”, reconstructed data etc. The same scheme can be applied to data from the associated sciences, with some differences regarding the tasks of data filtering and reconstruction.

Control data comprise but are not limited to calibration, positioning and conditions data which are accumulated and stored separately. These may be:

1. Detector control system data
2. Data quality/monitoring information
3. Detector and DAQ configuration information
4. Calibration and positioning information
5. Environmental data such as sea current velocity, sediment densities etc.

The diverse nature of the associated sciences measurements calls for a high degree of flexibility in the organization to accommodate the data of a multidisciplinary underwater observatory.

4. Data Management Services

A data management system is the basic infrastructure and tools that shall allow the KM3NeT institutes and physicists to locate, access and transfer the various forms of data in a distributed computing environment.

A typical Data Management System consists of the following components:

1. A Dataset Bookkeeping System – Which data exist
2. A Data Location Service – Where the data are located
3. A Data Placement and Transfer System
4. The Local File Catalogues
5. The Data Access and Storage Systems

A Storage System, typically referred to as Mass Storage System of MSS, is equipped with a Storage Resource Manager (SRM) interface providing an implementation independent way to access the MSS. Thus, while the database technology may change, applications accessing it will require minimal, if any, modification in order to continue functioning. Needless to say, such applications are obviously instrumented with I/O facilities that are user (physicist)-friendly. The File Transfer Service (FTS) must of course be scalable to the required bandwidth. The File Transfer Service implementations will depend on whether we opt for the use of a data GRID and the particular system we adopt.
A back-up route, should we experience port or other failures, would ensure the continued transfer of critical data and avoid backlogs and unacceptable processing delays.

All components are equipped with authentication, authorization and audit/accounting facilities. Such facilities ensure controlled access to the data (crucial during installation and early analysis efforts, as well as the tracking and logging of I/O operations and part of the data provenance.

Depending on the extent to which we choose to adopt GRID services and the implementation of choice, such facilities may be coupled to GRID utilities and tokens.

5. Database considerations

A Mass Storage System implies the use of one or more database technologies. Evidently database services must be based on scalable and reliable hardware and software.

For the latter, we may consider adopting packages and tools already in use in High Energy Physics, such as ROOT for event data and object/relational systems such as Oracle and MySQL for control data.

ROOT has proven reliable, flexible and scalable; it comes with a command line interface and a rich Graphical User Interface (GUI) as well as an I/O system, a parallel running facility and a GRID interface. It is easy to learn for users and developers alike. Its long-term support and maintenance are guaranteed by the ROOT team (ROOT is an LHC Computing Grid [2] Project at CERN and Fermilab).

ORACLE is the de-facto relational database standard whereas MySQL and PostGreSQL are open source, hence free, and may be adopted if cost concerns are prohibitive. However interoperability issues might pose problems and must be evaluated.

6. Discussion

Database options as well as various available implementations for the data management system components and services, regardless of the final computing model to be adopted, may need to be evaluated under conditions as realistic as possible. Possibly viable Mass Storage System implementations may include CASTOR used at CERN and dCache used at Fermilab and DESY. It is quite evident that some level of GRID use will be indispensable for the data transfer and distributed analysis "exercises" should be carried out.

Another important issue is that of (mock) data challenges. These can only be carried out once the basic elements of a structured system are in place and a large part of the necessary software for event simulation, reconstruction and analysis is commissioned and validated. However requirements as to scope and scale must already be collected and analyzed.

Acknowledgements

This work is supported through the EU-funded FP6 KM3NeT Design Study Contract N$^{o.}$ 011937.

References

[1] KM3NeT Collaboration, KM3NeT, Conceptual Design Report for a Very Large Volume Neutrino Telescope in the Mediterranean Sea, April 2008.
[2] LHC Computing Grid Technical Design Report, LCG-TDR-001, CERN-LHCC-2005-024, 20 June 2005.