Data Access and Management: Moving Beyond the MONARC Model

Introduction:
The majority of the LHC Computing Models are based to greater and lesser extents on the MONARC model (Models of Networked Analysis at Regional Centers for the LHC Experiments). The 4 LHC experiments have chosen to implement architectures based on elements from MONARC, but vary based on what elements, services, and facilities were emphasized. Assumptions and philosophies in the MONARC model have served the community well, but the MONARC hierarchies and fundamental assumptions are now more than a decade old and there have been significant advances in the field of distributed computing.

Many of the aspects of the LHC Computing models are products of compromise. The services provided by the grid function at a high scale and have been deployed on more than 100 sites worldwide, but the transparency of access and resource utilization promised in the earliest grid planning has never materialized. For most of the LHC experiments jobs are sent to data that has been previously placed at a site; data is either placed in disk copies and tape copies, or staged in an organized way from tape before processing; and generally the experiments have instrumented the workflows and applications with knowledge of what data to expect on the local site and how to access it. Many of the hierarchical elements of the original MONARC model are visible to the experiments and the users. The original hope for the transparent flow of data and requests to available resources has been traded for a more structured environment for many of the LHC experiments. Those that have provided remote access provide interesting examples to look at functionality and scale moving forward.

The structure of the existing system lends itself reasonably well to structured processing, and the organized processing systems are performing and scaling as expected. It is less clear that the more chaotic environment of user analysis with large LHC datasets will perform as well. The goal of the workshop is to evolve the existing implementations past the MONARC origins and to improve both the accessibility and manageability of data to support data in the future of the LHC.

Current Data Access and Management:

In the area of data access currently LHC applications are either instrumented with a lot of information regarding the local storage serving data or rely on a consistent data access infrastructure at all sites supporting the experiment. In order to achieve reasonable efficiency significant information regarding the local configuration needs to be available to the application or the environment needs to be consistent. Applications are configured for particular local protocols and optimized for the features of each protocol or the protocol is controlled. Applications that provide non-local access to storage have implemented large-scale central catalogs to provide centralized access control and applications that only access local data control the data available at the site. The experiment frameworks are configured to know the data available on disk and tape, staging data from tape to disk when necessary. The experiments have CPU efficiencies that match the expectations from the computing models, but significant optimization is being applied and it's not clear that as the access increases that the CPU efficiency will remain as high. Experiences with user analysis already show that the CPU efficiency is decreased as the access to data is more intensive.

In the area of data management the experiments allocate data globally by distributing individual files to local sites. Data is generally pre-placed for later processing by CPUs. Pre-placing the data leads to disk requirements that exceed the total volume of data collected, sometimes by integer factors, because multiple copies are needed to ensure the data is accessible to local processing resources. Providing remote access may reduce the disk requirement, but even the experiment that has this element in their computing model does not have sufficient pledged disk. The experiments currently without this functionality are finding the expected increase in LHC data volume will exceed the disk that can reasonably be provided without improvements in the architecture. Data management systems need to be improved to provide scalable analysis infrastructure as the volume of data being accessed increases.

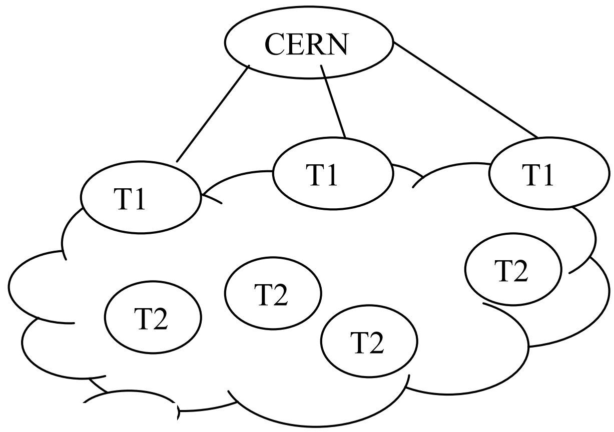
Technology Improvements:

A number of technology areas have improved dramatically over the development of the LHC.

1.) Networking: Once seen as one of the weakest elements of the computing models, the networks have proven themselves to be robust and networking technology to provide redundant paths and failover has improved the situation further. The connection from major research backbones to individual sites has improved dramatically over the last several years. There is always the need for additional capacity and the concern of oversubscription, but new tools for dynamic provisioning of the network can allow efficient allocation and predictable bandwidth.

2.) Data serving and peer-to-peer: There have been recent improvements in data streaming and data serving over wide areas both in the scientific and commercial communities. Combined with high capacity networking between sites, this may alleviate the need for some local storage by providing access to remote storage

Straw man proposal:
The LHC experiments should continue improving data transparency and data access efficiency. Specifically the effort should concentrate on reducing the dependency of reading data from tape, providing greater data transparency through a more advanced data access layer, and improving data management by automating more of the data placement and allowing remote access to data. The end result would be closer to a cloud model of data access and resource utilization than the MONARC model of data placement and site responsibilities.



The experiments should rely on the network and remote access to data more with the goal of improving access for analysis workflows and to reduce the total amount of disk necessary to support analysis.   Additionally, tape archives should evolve to be true data archives with facilities using data cached on disk, even if that data is pulled from a remote disk cache.   Reading from tape should be optimized with a number of other options including pulling from a disk copy on a remote site.

In the area of data access, the interface layer between the application and the local storage should evolve to allow the application to know less of the underlying storage and less about data locality.   Streaming some files over the WAN, or streaming the file over the WAN to local storage should become a standard.

In the area of data management, the remote access and improved data accessibility should be employed to allow LHC applications to operate efficiency with incomplete datasets hosted on the local site.   The remaining files could be streamed over the WAN or downloaded dynamically.   Peer-to-peer data serving and intelligent data placement should be investigated.

Timescale:
The timescale examined in this workshop is deployable tools and techniques in time for the 2013 run of the LHC.   This is medium term goal requires short-term progress on refining requirements, short term demonstration prototypes, and integration testing. Given the exciting 2010 and 2011 runs at the LHC it is of particular importance that the development of new technology not distract from the successful operations of the current system.