Evolution of WLCG Data & Storage Management

Summary of a first discussion on the strategy for evolution of the data and storage management in WLCG

Present:

ALICE (F. Carminati, L. Betev), ATLAS (K. Bos, G. Stewart, S. Campana), CMS (I. Fisk), LHCb (M. Cattaneo, A. Cameron); CERN IT (D. Foster, A. Pace, D. Duellmann) & WLCG (I. Bird, J. Shiers, M. Schulz, B. Panzer, M. Girone).

Introduction

The LHC experiment managements have expressed concern over the performance and scalability of access to data, particularly for their analysis use cases. This meeting, held to provide a first response to these concerns, focused on setting the scope and goals for work that that would address these issues with a tentative timescale of 2013 for large scale use. It is anticipated that following this meeting a series of working groups to address particular technical areas would be set up, with the goal of producing incremental working prototypes to validate some of the ideas. These working prototypes should be of immediate use and help to resolve some of the shorter term concerns of performance and functionality. There must be incremental improvements to the current system. A "jamboree" workshop will be held to review the available tools and technologies, and to elaborate implementation plans. These meetings and follow-up workshops would aim to be as inclusive as possible, whilst remaining focused on the issues at hand. It is important to recognise that this work should wherever possible make use of existing tools and software and not be seen as a rationale for new development projects. It is essential that new services and tools are supportable and sustainable in the future and do not rely on ongoing development projects.

This work must start from the viewpoint of user access to data and the resulting system should hide the details of the back end mass storage systems and their implementations. The outcome of this work must be driven by the needs of the experiments and their use cases. In the discussion this was focussed on analysis use-cases as the organised production work, which in future may benefit from developments in this area, today work sufficiently well not to be an immediate cause for concern.

As a basis of the discussions it was recognised that there have been significant technology developments since the MONARC model was first proposed and that we must take advantage of them. These include: networking, where the available capacities and reliabilities are not fully utilised in today's computing models; available aggregate disk capacities far in excess of what was originally anticipated in the early models; and other interesting developments such as virtualisation and large scale file systems.

Note: It is not the intent to replace the working system that is currently in place during the first years of data taking, but to prototype solutions for the future, with a tentative timescale of 2013 for large scale use.

Areas of work

A working model is to be far more network-centric idea than today. The initial hypothesis is to have a (few) large archival data repositories that are responsible for long term data curation, and a cloud of storage used as short term data caches with peer-peer technologies used to transfer data. Data may be accessed from the cache or across the network. A common data access layer provides some intelligence in optimising this. Specific areas where work is needed are the following:

1. Data Archives and Storage Cloud. A simplified model where the "tape" back ends are treated as truly archival storage (i.e. data to be read only when cached data is corrupted or

by managed stage out by experiment coordinators). The interface to archival storage then becomes very simple – essentially put and get. The disk storage at Tier 1 and 2 sites then should be seen as a cloud of disk caches with the assumption that data can be actively moved (and cached) so that work can run without the requirement that all the data that it needs be located in one place. Jobs should be able to request data be accessed remotely, either by a local cached copy being made or by reading the data remotely. The details of this must be transparent to the user. Such a model can potentially make better use of available resources, but may require additional investment into networking capacity. New transfer technologies including peer-to-peer should be investigated as solutions in this area.

2. Data Access Layer. Here the working hypothesis would be something like xrootd/GFAL, possibly with some additional intelligence to understand where files are, when to cache and when to use remote access. Robustness of this interface to file availability is a key concern here.
3. Output Datasets. Datasets created by analysis jobs or simulation tasks will still require a service to (asynchronously) migrate them to a data archive, which may require a point to point asynchronous transfer service.
4. Global home directory facilities. This is an important missing functionality today. Global file systems are the model of what is wanted, and may provide the solution. Industrial solutions are clearly preferable.
5. Catalogues. This still a need – the knowledge of the location of data is still required. The issue of consistency between storage system catalogues and "grid" catalogues must be addressed (perhaps by removing the need for different catalogues).
6. Authorization mechanisms for access to files in storage systems (archive and cloud), and quotas are required.

Next steps

1. Jamboree – 3-day workshop to look at existing tools in each of these areas and how they could be used. Early June, location to be agreed, ~100 people?
2. Following the Jamboree: Elaboration of a more concrete plan and timelines for each of the above technical areas. Set up of working groups.
3. Develop demonstrator prototypes in each area – testing of individual components/technologies
4. Experiment testing – integration of solutions into the experiment frameworks