September, 2010

Network Requirements

Introduction

When in the late nineties the computing models for the 4 LHC experiments were developed, networking was still a scarce resource and almost all models reflect this - although in different ways. Instead of relying on the ability to provide the data when needed for analysis, the data is replicated to many places on the grid shortly after it has been produced to be readily available for user analysis. This model has proven to work well for the early stages of analysis but is limited by the ever-increasing need for disk space when the data volume from the machine increases with time.

Currently, the network is the most reliable resource for LHC computing. This makes it possible to reconsider the data models and not rely on pre-placement of the data for analysis and running the jobs only where the data is. Instead, jobs could pull the data from somewhere else if not already available locally. It will depend on the data needs to decide whether to copy the data locally or to access the data remotely. If the data is copied locally the storage turns into a cache that is likely to hold the selection of the data that is most popular for analysis at that time. This solves the problem of the current models where, at the time of the placement, it is not known which data will be wanted most - and hence all data will be placed indiscriminately.

There were two workshops conducted in June of 2010 that provide input and ideas for this document. The LHCOPN conducted a workshop at CERN on Transatlantic Networking in June to discuss how to improve the connectivity between LHC Tier-2s. Additionally, the WLCG and the LHC experiments hosted a workshop on Data Access and Management in Amsterdam. One of the main themes of the workshop was improving the use of the network to facilitate data analysis.

As analysis is mainly done in Tier-2s, it is likely that if data is missing for a particular task the data can best be copied from another Tier-2. The current ATLAS data-model data can only be pulled to a Tier-2 from the Tier-1 of the same cloud. The CMS model is more advanced and Tier-2s can pull data from any Tier-1. Ideally, data should be able to be pulled from any Tier-1 but, more importantly, also from any Tier-2 – or even, if possible, from several Tier-2s simultaneously in order to minimize the time to have the data locally to run the analysis task.

The LHCOPN provides the infrastructure to efficiently move data from the Tier-0 to the Tier-1s and between the

Tier-1s. Within the ATLAS clouds - and for CMS across clouds - data can also be transferred from the Tier-1 to the Tier-2s. A new infrastructure is now needed to make it possible to improve the transfers between Tier-1s and Tier-2s and to make efficient Tier-2 to Tier-2 possible. This note collects requirements for this new infrastructure.

Responsibilities

It is the experiments' responsibility together with the Tier-2 sites to describe the model for data caching and analysis. The formulation of those requirements such that they allow choosing between appropriate network architectures is the responsibility of the network specialists in this group. The requirements document needs to be approved within the experiments and by the CERN(/IT) management. The LHCOPN group will need to decide and describe which is the best network architecture that will fit the requirements.

A Tier-2 Model

Tier-2s are used primarily for detector simulations and for analysis. Typically, 50% of the capacity is used for each, although in periods when there is little demand for simulation all capacity is used for analysis and vice-versa. Simulation is very CPU demanding and has very little demand on the network but analysis is the opposite. Grosso modo, one can therefore assume that the network requirements of a site are driven by 50% of its CPU capacity.

The network needs of a Tier-2 can be motivated from a variety of perspectives: the storage, the user, and the processing. From a storage perspective, a nominal Tier-2 in either ATLAS or CMS has a few hundred terabytes of disk space. During large scale reprocessing efforts, a significant portion of the storage data will need to be refreshed. To refresh a 400TB disk cache in a week requires access to 5GB/s of incoming network.

Given the experience with the first year of data collection we can look at the datasets accessed by users and look at a reasonable latency to refresh a user sample. Unlike the large scale data refresh that will happen at predictable intervals, updating user samples will have bursts driven by user requests. In 2010, analysis users are accessing samples over a large range of sizes. A few terabytes is common, but even with the limited statistics collected so far samples in the 10s of terabytes are used for analysis. To transfer a 25TB sample in 24 hours requires 3Gb/s continuously.

One can also look at the bandwidth needed simply to serve data to the Tier-2 analysis CPUs and estimate the rate for 1000 cores. Given the current event size and application speed in ATLAS and CMS, this results in

approximately 1Gb/s - which is, in some sense, the average rate needed and does not include the bursts.

The three approaches give a range of values but provisioning factors also need to be included. Given the spread, it makes sense to attempt to categorize the Tier-2s connections.

Category	Speed	Target
Minimal	1Gb/s	Small Tier-2 installations: at the minimal connection speed a Tier-2 will be able to function, but will not be able to provide users with the same flexibility and quality of service.
Nominal	5Gb/s	Normal Sized Tier-2 installations: at nominal connection speeds the samples can be updated in reasonable time and the Tier-2 storage can be updated at regular intervals.
Leadership	10Gb/s and greater	Large Tier-2 installations: leadership Tier-2 facilities are significant analysis facilities supporting large numbers of analysis users. The high connection speed allows the large local storage to be updated and samples provided to several individual users working simultaneously.

It must be stressed that this is an oversimplified model, and it is easy to find arguments that can influence its results upwards and downwards - but it gives an indication of the typical bandwidths needed for the three classes.

Looking forward, there are a number of factors that influence the bandwidth needs into the future. LHC data analysis is still in early days and the data volume is still expected to grow significantly. The processing capacity and subsequent analysis capability at Tier-2s will continue to increase with computing improvements. At the same time, the number of users will grow, but probably not more than a factor 2. The minimal network category will likely grow by a factor of 2 every year as data volume increases. The nominal and leadership categories should double every two years.

In order for the Tier-2s to reach their analysis potential, the site network connectivity needs to allow them to communicate with the Tier-1s and the other Tier-2s. The bandwidth needs are driven by data updates and bursts from analyzers, so the backbone does not need to support all possible connections at full speed all the time. The backbone does need to support several full

speed connections between the leadership Tier-2s simultaneously.

Other than Bandwidth Requirements

Staging

In both ATLAS and CMS, there are sites that currently play a leading role in analysis and connecting them with better networking will have a larger effect in facilitating this. 75% of ATLAS analysis is done in 25% of the sites. In CMS, the usage is somewhat flatter - with 75% performed at 35% of the Tier-2 centers. Doing analysis efficiently is not an easy task and, in general, those Tier-2s are well staffed with dedicated experiment people and with well-installed and maintained hardware. For ATLAS the other Tier-2s that don't meet these standards (yet) are primarily used for simulation production. In CMS only the smallest Tier-2s are used exclusively for simulation. It is in the interest of both experiments that the Tier-2s with the highest impact in analysis are well connected.

However, sites that are currently used for simulation only may become eligible for analysis at some later time. Moreover, caching data for analysis serves better the smaller sites that are currently not used because their disk space is not sufficient for data pre-placement. Were they to be given sufficient network connectivity to cache smaller subsets of data regularly, they could participate in analysis even though they don't have the disk capacity to hold all the data (which is at present the requirement).

Both ATLAS and CMS have large analysis centers in Europe, North America, and Asia. With 34 countries participating in the LHC program, there are many sites in countries that need additional effort to connect. South East Asia, India and China, Africa and South America all have active analysis facilities with particular network challenges.

The attached table ranks the Tier-1s and Tier-2s by size in processing capacity. The size of the centers is a strong indicator of their relative contribution to analysis activities and improving the connectivity of the largest centers will have the greatest impact at the beginning.

Flexibility

The Tier-1s and the most important analysis Tier-2s can be assumed to comprise a fairly stable collection of sites. The architecture can take this into account but has on the other hand to be very flexible because of the bigger number of Tier-2s in the other category. This may be a more rapidly changing collection where sites disappear for political reasons or they loose their

funding and other sites will appear for opposite reasons. Those sites are from a bigger variety of countries than the Tier-1s were, and different National policies may have their impact.

Budget neutral

Additional costs for networking have not been included in the budget estimates for 2011 until 2013. Although data caching may initially reduce the ever-increasing disk space needs, in the long run more disk space will always be needed. Some of the analysis Tier-2s recently had to invest in networking equipment local to the site to achieve the rates mentioned above between the storage and the CPU. Those sites will more easily understand that a better connection to the wide area will need investments. For the smaller Tier-2 sites that were so far served by their campus network and the public internet, this may not be so obvious.

Sites Ranked by Size

Facility	Tier
FZK- GridKa	Tier-1
IN2P3	Tier-1
Netherlands	Tier-1
INFN CNAF	Tier-1
US-FNAL	Tier-1
UK	Tier-1
US-BNL	Tier-1
Italy-CMS	Tier-2
GRIF	Tier-2
Taipei	Tier-1
Russia	Tier-2
UK, London	Tier-2
IN2P3	Tier-2
Spain	Tier-1
Italy-ALICE	Tier-2
Romania	Tier-2
NDGF	Tier-1
Italy-ATLAS	Tier-2
Estonia	Tier-2
Spain-CMS	Tier-2
UK, SouthGrid	Tier-2
USA-CMS-Caltech	Tier-2
USA-CMS-Florida	Tier-2
USA-CMS-MIT	Tier-2
USA-CMS-Nebraska	Tier-2
USA-CMS-Purdue	Tier-2
USA-CMS-UCSD	Tier-2
USA-CMS-Wisconsin	Tier-2
UK, ScotGrid	Tier-2
Spain-ATLAS	Tier-2
Switzerland-CHIPP	Tier-2
Poland	Tier-2

USA-ATLAS-NE	Tier-2
USA-ATLAS-SW	Tier-2
USA-ATLAS-Midwest	Tier-2
USA-ATLAS-Great Lakes	Tier-2
USA-ATLAS-SLAC	Tier-2
Tokyo	Tier-2
CMS-DESY	Tier-2
Canada	Tier-1
UK, NorthGrid	Tier-2
Mumbai	Tier-2
Turkey	Tier-2
Belgium	Tier-2
ATLAS Munich	Tier-2
Strasbourg	Tier-2
Czech	Tier-2
China	Tier-2
Slovenia	Tier-2
Sweden	Tier-2
Canada-East	Tier-2
Canada-West	Tier-2
CMS-RWTH	Tier-2
LPC	Tier-2
Portugal	Tier-2
ATLAS-DESY	Tier-2
Italy-LHCb	Tier-2
GSI	Tier-2
Austria	Tier-2
LAPP	Tier-2
Kolkata	Tier-2
ATLAS-Wuppertal	Tier-2
ATLAS-Freiburg	Tier-2
Taipei	Tier-2
Pakistan	Tier-2
Austraiia	Tier-2
Israel	Tier-2
ATLAS-Goettingen	Tier-2
IN2P3-CPPM	Tier-2
Hungary	Tier-2
LHCb-DESY	Tier-2
Korea-KNU	Tier-2
Nantes	Tier-2
Norway	Tier-2
Finland	Tier-2
Brazil	Tier-2
Spain-LHCb	Tier-2
Ukraine	Tier-2
Korea-Daejaon	Tier-2

The list is ranked by size, with predominantly Tier-1s at the top. These facilities are large and will become a source of data for all Tier-2s. Additionally, many Tier-1 centers have co-located Tier-2 centers for analysis. Next in the list are the largest Tier-2s.

All the centers in the first half of the list would be categorized as leadership Tier-2 centers.