Subject: Fwd: summary of the provided material and request for additional input
From: Markus <markus.schulz@cern.ch>
Date: Tue, 17 Mar 2009 16:04:14 +0100
To: "worldwide-lcg-management-board (LCG Management Board)" <worldwide-lcg-management-board@cern.ch>

I have sent the following summary and set of questions to the Analysis Working group to move forward.

markus

Begin forwarded message:

> From: Markus <markus.schulz@cern.ch>
> Date: March 17, 2009 3:56:35 PM CEST
> To: <project-wlcg-uag@cern.ch>
> Cc: <project-wlcg-user-analysis-wg@cern.ch>
> Subject: summary of the provided material and request for additional input
>
> Dear User Analysis Working Group,
> here a summary of the provided material.
> At the end of the summary you can find a list of open questions and topics that need more quantitative input.
> In addition a first list of requirements that have been mentioned frequently and are analysis specific.
> Please comment and provide input where you can.
>
> markus
>
> p.s. I attached slides that I have shown a few weeks ago to the LCG-MB.
>
> Based on the input received from the four experiments I try to summarize the situation.
>
> All experiments have gained significant experience with their user analysis frameworks and workflows, which are well described in their functionality.
> It is not clear from the provided material to what extend the scale of the current experience reflects the expected scale. The feedback from endusers of all experiments are similar.
> Individual analysis campaigns start with high failure rates in the order of 30%. All experiments put monitoring and communication systems in place to validate sites for suitability for analysis
> and work directly with the site managers on resolving local issues and tuning the resources.
>
> When it comes to the number of active users and the number of jobs per user per available CPU unit the material provided doesn't provide clear information.
> Currently in dashboards indicate that the number of different
> concurrently active users for all experiments is about 300-500 during a week and about 1500 distinct users. It is not clear how this translates into the number of individual
> users that will perform analysis tasks on the LCG grid infrastructure.
>
> CMS has been very clear on the requested resource split on T2s between the analysis and reconstruction computing resource share. For other experiments it is more difficult to get a clear
> picture from the provided material.
>
> The approaches to analysis differ between the experiments significantly.
> High level descriptions and block diagrams suggest some similarities, however they vary greatly when it comes to implementation.
>
> These differences result into dependencies of the experiments on different components of the grid infrastructure and different efficiencies for basic operations, such as file I/O.

The recent studies by Andreas Peters, Max Baak and Matthias Schott are good examples to illustrate these differences.
Using the same underlying software and local files the I/O performance between two experiments and the ROOT example cover a range from 2.2 - 47 MB/s.
Furthermore there is a strong dependency between the storage system implementation and the efficient usage of the resource. Experiments such as ATLAS react dynamical on the situation and adapt their data access mechanism to sites based on experience. ALICE on the other hand strives for accessing data only via xrootd either directly or through xrootd interfaces to SRM based storage.

Each experiment follows a different approach for access to computing resources for user analysis.
ALICE and LHCb base their access on pilot frame works, ATLAS and CMS use and plan to use pilot based access and the WMS. In addition interactive analysis outside the grid infrastructure via PROOF is important for ALICE.

With all the differences there are some requirements that are shared between the experiments.
Except of requests related to better availability and reliability one set of requests is related to authorization and the control of resource usage, the other to the scalability of core services such as the SRM interface, the CE and the LFC. Not all of these requirements are user analysis specific.

It should be noted that T2s, where most of the analysis will take place, show much more variation in size than the T1s. The largest T2 currently provides more than 6000 cores and there are several with less than 100 CPU cores. It is unlikely that there will be recommendations that apply to all of them in the same way.

Functional computing resource requirements:

0) Support for multi user pilot jobs

1) The ability to assign shares and priorities to different analysis group taking into account the locality of a user.

2) Fair share allocations have to balance within predefined time windows.

3) Prioritization for individual users based on recent usage.

0) Is addressed on the OSG sites via GUMS/glexec, SCAS/glexec can address this on EGEE sites used by LCG.
For NDGF, with the approach to not require middleware on the WN this needs clarification.

1) The current system based on VOMS groups and roles and the VOviews system on the CE and in the information system can be used to express the desired behavior.
However, with multiple analysis groups and other tasks hosted at a site the number of VOviews that need to be handled is potentially large and will effect the information system.
Here some clarification is needed to understand to what extend higher shares for local users are better implemented via configuration of the batch system scheduler or via the authorization on the CE for specific shares. This is best discussed between T2 admins and batch system experts.
To progress on this input from the experiments is needed on how many different analysis working groups they expect on a site .

Requirements 2 and 3 can be addressed on most sites by changes to the configuration of the local scheduler. Users noted that most sites do not show the desired behavior and suspect that this is related to a lack of in depth understanding of the configuration. This is best addressed by providing an improved configuration guide.
Related to batch system configuration is the problem that the lcgadmin roles are not always mapped to the right local accounts and priorities.

Alice expressed clearly that they would like to see more Cream-CE instances in

> production. This is not directly related to analysis.
>
> Scalability issues related to computing resource access:
>
> To give the T2s an indication on how many CEs they have to deploy the expected scale of the workflow on a T2 is needed. The current production CE (LCG-CE) on EGEE cannot handle a large number of concurrent group, role and user combinations well.
> The limit is about 50 role/user combinations while handling 4000 jobs.
> For this the T2s need to know the number of expected users and analysis groups per unit of CPU.
>
> Storage Requirements
> =====================
>
> CMS provided a very useful table on the required disk space per analysis group and users on a T2.
>
> \>20TB temporary space for production, controlled by prod team
> 30TB space for centrally managed official data sets
> N* 30TB for each official physics group
> Regional (local) user spaces
>
> Functional Requirements related to Storage
> ====================================
>
> 0) SRM APIs and tools, especially to handle bulk operations and frequently executed commands such as "ls" .
>
> 1) ACLs (VOMS based) on spaces and files.
>
> 2) Quotas
>
> 3) Accounting
>
> From the material provided it is not entirely clear on what granularity quotas are required. This needs more clarification. DPM and dCache have already support for access control based on VOMS extensions.
>
> As part of the discussion on quotas and accounting for Castor some questions have been raised concerning the detailed expected behavior.
> Here a reminder of some of
> Dirk's comments from last year:
>
> > a) system provide per user disk storage accounting
> >
> > define a volume metric and provide a summary per storage area (eg few times per day and pool)
> >
> > b) user based quota
> >
> > system knows about individual (or common?) volume quota values and follows some escalation policy (warning email, writes fail, etc..), who is in the reply field of the warning/blocking email?
> >
> > \- this would mean that quota admin tools exists and procedure for deployment / experiments responsible to change quota are existing
> >
> > \- does quota ever shrink? if yes, how?
> >
> > \- confirm that the user IDs used currently the ones relevant for the accounting/blocking
> >
> > \- can we ignore/directly use the users coming from any of the grip mapping steps
> >
> > c) group based quota

> - based on what groups (LSF, VOMS, other adhoc groups)
>
> - would require tools to sync / define groups, procedures for deployment/experiment responsible to change/admin those groups
>
> - else, should/could this be done based on user accounting inside experiment?
>
> a'-c') name space based storage allocation?
>
> - may be an alternative to user based accounting/quota
>
> - basically as above but accounting is done per pool and data amount below some directory in the name space
>
> - escalation as above to name space "owner" (which could be a user for their home dir or a calibration group for their shared area)
>
> - may not be feasible, but may require less additional tools to administer and sync with other user and group management systems

Storage Access

There has been the requirement by ALICE to have access to all storage via xrootd.

Given the interesting results by Andreas, Max and Matthias it might be worth to carry out similar tests for all experiments to understand the load on the storage systems and networks on T2s.

Storage management tools

In some of the provided material LFC and bulk registration was mentioned. However was not clear to what degree this is directly related to user analysis.

Open questions to be discussed:

Expected number of active users, users per T1/T2/T3, this can't be a fixed number, what is the ratio between CPU/user? How many jobs per day and user do you expect? Peak, typical?
------> Experiments

How many different analysis working groups will be supported on a given T1,T2,T3?
-----> Experiments

Would special queues for pilot jobs be an advantage to organize the workflow for experiments that use both approaches? What characteristics are desirable (job duration etc.)? ---> Experiments

Batch system reference setup and improved scheduler configuration guide. -----> T2 admins + Steven Traylen.

Expected fraction of analysis via analysis train, independent user analysis and proof. ---> ALICE

NDGF Support for multi user pilot jobs, how and when? ----> NDGF

What fraction at a given T2 do you expect to be available for analysis? ---> Experiments
What fraction of the analysis workload will flow through your pilot job framework?
-----> Experiments

Provide the required storage for analysis groups and supported users. ---> Experiments

Clarify the SRM API requirements and tools, there is already Flavia's team collecting input, a list of most needed APIs and the

> expected performance by experiment would definitely help. Some feedback from larger T2s would be appreciated too.   -----> Experiments and T2s and data management experts
>
> Storage quotas and accounting, what granularity is required, what can realistically been implemented for which storage system, to which degree are the items on Dirk's list now understood?   ----> Experiments and storage experts
>
> Storage access efficiency, ATLAS and ALICE tests (A,M,M-test) to be done for CMS and dLHCb to get numbers for usable MB/s and overheads. ------> Experiments and T2s and storage managers
>
> Bulk registration of files in LFC, are there other scalability issues related to file management tools? ---> experiments and data management experts

User_AnalysisWorkGroupUpdate.ppt.pps	**Content-Type:**	application/vnd.ms-powerpoint
	Content-Encoding:	base64

Part 1.3	**Content-Type:**	text/plain
	Content-Encoding:	7bit