Tier-1 Services for Tier-2 Regional Centres

The LHC Computing MoU is currently being elaborated by a dedicated Task Force. This will cover at least the services that Tier-0 (T0) and Tier-1 centres (T1) must provide to the LHC experiments. At the same time, the services that T1s should provide to Tier-2 centres (T2) should start to be identified and described. This note has been written by a small team appointed by the LCG PEB with the objective of producing a description of the T1 services required by T2 centres. The members of the team are:

Gonzalo Merino /PIC - convener
Slava Ilyin / SINP MSU
Milos Lokajicek /FZU
Klaus-Peter Mickel /FZK
Mike Vetterli / Simon Fraser University and TRIUMF

The T2 requirements on T1 centres identified in this note have been mostly extracted from the current versions of the computing models of the LHC experiments. These are still in active development within each of the experiments. Therefore, these requirements will need to be revised as the computing models evolve.

Experiments' computing models plan for T1 and T2 centres:

- Tier-1:
  - To keep certain portions of RAW, ESD, simulated ESD data and full copies of AOD and TAG data, calibration data.
  - Data processing and further reprocessing passes.
  - Official physics group large scale data analysis (collaboration endorsed massive processing).
  - ALICE and LHCb – contribution to simulations.
- Tier-2:
  - To keep certain portions of AOD and full copies of TAG for both real and simulated data (LHCb – store only simulated data at T2s).
  - To keep small selected samples of ESD.
  - Produce simulated data.
  - General end-user analysis.

The T2 requirements on T1s identified in this document emerge from these roles and their interplay. They have been categorized in five groups and each of them is described in one of the following sections.

1. Storage Requirements

There is a wide variation in the size of T2 centres. Some will have a significant fraction of the resources of a T1 centre, while others will simply be shared university computing facilities. The role of the T2s even varies from experiment to experiment. This makes it somewhat difficult to define a standard set of requirements for T2s. Nevertheless, the

following describes the services that T2s will require from T1s with regard to storage. These are listed in no particular order of importance.

1) Some analyses based on AODs will be done at the T2s. The T1s will therefore need to supply the AODs to the T2s. This should be done within 1-2 days for the initial mass distribution, but the timescale should be minutes for requests of single files in the case that the T2 centre does not have the AOD file required by the user. In the latter case, the missing AOD file could also be downloaded from another T2 center.
2) During the analysis of AODs, it is possible that the T2 process will need to refer back to the ESDs. A subset of the ESDs will be stored at the T2s but it is likely that the particular data needed for analysis will be at the T1. Access to single ESD files at the T1s from the T2s should be on the timescale of minutes.

These first two points will require that access to the data files stored at the T1s be Grid-enabled so that the process of location and retrieval of data will be transparent to the user.

3) The T2s will need to store a subset of the raw data and the ESDs for algorithm and code development. They will get these files from the T1s.
4) One of the identifiable roles of the T2s is Monte Carlo production. While T2 centres are likely to have the CPU power necessary for this task, it is unlikely that sufficient storage will be available. The T1s should therefore be prepared to store the raw data, ESDs, and AODs from the Monte Carlo production. For ATLAS, this corresponds to 200 TBytes for the raw data, 50 TBytes for the ESDs, and 10 TBytes for AODs per year. Since the ESDs will be replicated twice across all T1s and each T1 will store the full AOD, this leads to a total of 360 TB per year spread across all T1 centres for ATLAS Monte Carlo. This requirement will be even larger if multiple versions of the ESDs and AODs are produced each year. CMS plans to produce an equivalent amount of Monte Carlo data to real data so that CMS T2s will require as much storage at their corresponding T1s as for real data. The number for LHCb is 413 TB of Monte Carlo data per year, augmented by whatever replication factor is applicable for LHCb. The total storage for Monte Carlo data at ALICE is 750 TB/year, but this will be split equally between the T1 and T2 centers (with a small amount, 8%, at CERN).

The large file transfers of Monte Carlo data from the T2s to the T1 mass storage systems (MSS) should be made as efficient as possible. This requires that, for example, the MSS should have an SRM interface[1].

5) The T2 centres will also need to get the calibration and slow controls databases from the T1s.
6) ALICE: The computing model at ALICE is somewhat different from ATLAS and CMS. T1 and T2 centers play essentially the same role in the analysis of the data. The main difference between the two is that T1s have significant mass storage and will therefore be responsible for archiving the data. ESDs and AOD analysis

[1] http://sdm.lbl.gov/srm-wg/

will be spread over all T1 and T2 centres, with 2.5 copies of the ESDs and 3 copies of the AODs replicated over all T1 and T2 centers.

7) The T2 centres will be heavily used for physics analyses based on AODs. The results of these analyses (e.g. ntuples) will need to be stored somewhere. Those T2s with mass storage can do this for themselves. However many T2s, especially those in university computer centres, will have mass storage only for backup of user home areas, not for data or large results files such as ntuples. In these cases, it will be necessary for the T1s to store the results of user analyses on tape. This could amount to about 40 TB per year per experiment; the numbers in the current models for CMS and ATLAS are 40 TB and 36 TB respectively.

2. Computing Power Requirements

The T2 centres will have no special requirements for usage of T1 CPU resources. The CPU use will be primarily the decision of experiments on resource allocation for specific tasks. The Data Challenge results will influence experiments computing models towards usage of T1 and T2 centres. ALICE keeps its model flexible on the load distribution between the centres. According to current computing models the T1 centers except CERN resources deliver 50% of computing power for ATLAS and CMS, 40% for ALICE and 30% for LHCb experiments.

General assumption is that processing of data should be preferentially done at centres where the data reside. Certain amount of the T1s CPU cycles will be needed for data transfers (both remote data access and file transfer) to and from T2 centres. The transfer should not influence T1s computing elements power as it would be delivered by storage elements that will have to be balanced in respect to CPU power, disk space, data transfer requests and transfer rates.

One exception would be if the user analysis task processing AOD file would require access to missing information located on ESD/RAW data files. The requested file might be either remotely accessed, transferred to T2 or a remote task could be initiated on the T1 to process requested information. In the last, probably rare case, T1 CPU cycles would be required for T2 analysis tasks, but no estimates are available. Alternative solution is to process the tasks requiring ESD/RAW data on the T1 possibly as part of already mentioned large scale data analysis.

Computing models can enable the usage of the T1 centres free capacity by tasks normally processed at T2 centres like simulations or physicists analysis. For such usage T1 centres should provide:

- Grid enabled CPU cycles. Resource brokers of the experiment must be able to send jobs to the T1 resources and, from these jobs, access any grid enabled file.
- Possibly advanced reservation of CPU resources.

Computing models anticipate the task distribution between the T1s and T2s and thus usage of available CPU power. A substantial hidden need of T1 CPU power for the T2s

was not found once the conditions for users' T2 analysis tasks needing ESD information or usage of the T1s free CPU cycles has been covered in the computing models.

3. Network Requirements

The activities foreseen in the T2 centres are mainly Monte Carlo production and end-user data analysis. Therefore, in order to estimate the network bandwidth needed between a given T2 and its reference T1 centre, the following data transfer categories have been considered:

- Real data:
  - *From T1 into T2:* distribution of selected samples of RAW, ESD, AOD and TAG to T2s for further analysis.
- Monte Carlo:
  - *From T2 into T1:* copy of simulated data produced at the T2 into T1 for permanent storage there.
  - *From T1 into T2:* copy from the T1 the share of simulated data generated at other centres that should be available at the T2 for analysis there.

The numbers assumed here are those in the experiment computing models presented in the context of the group set up inside the LCG project to provide answers to questions posed by the Computing MoU Task Force[2]. A summary of the data in those models that is relevant for the T1 to T2 network services is presented in Table 1.

The bandwidth estimates have been computed assuming the data are transferred at a constant rate during the whole year. Therefore, these are to be taken as very rough estimates that at this level should be considered as lower limits on the required bandwidth. To obtain more realistic numbers, the time pattern of the transfers should be considered, but this is still very difficult to estimate today in a realistic manner. Furthermore, it is also very difficult to estimate the efficiency with which a given end-to-end network link can be used, given the number of factors that can affect that (fault-tolerance capacity, routing efficiency along individual paths, etc). In order to account for all these effects, some safety factors have been included. The numbers have been scaled up, first by a 50% factor to try to account for differences between "peak" and "sustained" data transfers, and second by a 100% factor in the assumption that network links should never run above their 50% capacity. The former would account, for instance, for the use case discussed in previous sections in which data replication from T1 to T2 is triggered by user analysis running at the T2 requiring access to some AOD/ESD that is not available at the T2. A substantial bandwidth should be "reserved" so that this replication could take place in a timescale of minutes.

[2] http://lcg.web.cern.ch/LCG/peb/mou

Table 1 - Bandwidth estimation for the T1 to T2 network links.

	ALICE	ATLAS	CMS	LHCb
Parameters:				
Number of Tier-1s	4	6	6	5
Number of Tier-2s	20	24	25	15
Real data "in-T2":				
TB/yr	120	124	257	0
Mbit/sec (rough)	31.9	32.9	68.5	0.0
Mbit/sec (w. safety factors)	95.8	98.6	205.5	0.0
MC "out-T2":				
TB/yr	14	13	136	19
Mbit/sec (rough)	3.7	3.4	36.3	5.1
Mbit/sec (w. safety factors)	11.2	10.2	108.9	15.3
MC "in-T2":				
TB/yr	28	18	0	0
Mbit/sec (rough)	7.5	4.9	0	0.0
Mbit/sec (w. safety factors)	22.5	14.7	0.0	0.0

The numbers that result from the computing models categorize the experiments in two different groups. On the one side there is LHCb, with the smallest bandwidth need estimated to be of the order of 15Mbit/sec. This is in part due to the fact that LHCb does not foresee to replicate to T2s any real data or Monte Carlo produced in other centres. On the other side, the estimated bandwidth needs for T2 centres in ATLAS, CMS and ALICE is of the order of 100-200Mbit/sec.

We want to stress at this point that the uncertainty in the safety factors assumed in this note is very large at this moment. For this reason, the numbers before and after applying such factors are quoted in the table. Specific tests should be performed during the experiments Data Challenges that address this issue. For instance, some recent experimental results from the current ALICE Data Challenge indicate that the network bandwidth needed between T1 and T2 centres could be as high as 100MB/sec.

The T1 and T2 centres located in Europe will be computing facilities connected to the National Research and Educational Networks (NRENs) which are in turn interconnected through GÉANT. Today, this infrastructure already provides connectivity at the level of the Gbit/sec to most of the European T1 centres. By the year the LHC starts, this network infrastructure should be providing this level of connectivity between T1 and T2 centres in Europe with no major problems.

For some sites in America and Asia the situation might be different, since the trans-Atlantic link will always be “thin” in terms of bandwidth as compared to the intra-continental connectivity. T1 centres in these countries might need to foresee increasing their storage capacity so that they can cache a larger share of the data, hence reducing their dependency on the inter-continental link. T2 centres will in general depend on a T1 in the same continent, so their interconnection by the time LHC starts should also be at the Gbit/sec level with no major problems.

According to the above numbers, this should be enough to cope with the data movement in ATLAS, CMS and LHCb T2 centres. On the other hand, those T2 centres supporting ALICE will need to have access to substantially larger bandwidth connections, since the estimated 100MB/sec would already fill most of a 1Gbit/sec link.

It is worth to noting as well that the impact of the network traffic with T2 centres will not be negligible for T1s as compared to the traffic between the T1 and the T0. The latter was recently estimated in a report from the LCG project to the MoU task force[3]. The numbers presented in this note indicate that, for a given T1, the traffic with a T2 could amount to ~10% of that with the T0. Taking into account the average number of T2 centres that will depend on a given T1 for each experiment, the overall traffic with T2s associated with a given T1 could reach about half of that with the T0. On the other hand, it should also be noted that the data traffic from T1 into T2 quoted here represents an upper limit for the data volume that a T1 has to deliver into a given T2, since most probably there will be T2-to-T2 replications that will lower the load on the T1.

4. Grid Services Requirements

Computing models of all four LHC experiments assume that T2 centres will operate as GRID sites in one of the (multi regional) infrastructures – EGEE in Europe[4], GRID3/OSG in USA (now the prototype is GRID3[5]), <Asia-LCG> etc., following the GRID-federated approach to the global structure of LCG. The main functions, to be provided by T2s (simulation and user analysis), tell us that they are *resource centres* (using the EGEE terminology). Then, core and operation services will be provided for them by corresponding GRID service centres, e.g. ROCs and CICs in EGEE (*Regional Operations Centres* and *Core Infrastructure Centres,* correspondingly). In the following we use the terminology *GRID Operation Center (GOC)*, as a generic name, e.g. for ROCs and CICs in EGEE.

In many cases GOCs will be hosted at T1s. Note, however, that in some regions GOC functions are distributed over several laboratories, most of them are T2s. One should add, however, that in these cases a single representative body should be defined for such distributed ROC and CIC. The T2 centre is treated by the "MoU for Collaboration in the Deployment and Exploitation of the T1 centres of the LCG" document (being under preparation still) as "a regional centre, the LCG-related work of which is coordinated by a defined T1 centre". In the following we refer on this defined T1 as to *hosting-T1*, while a T2 centre under this coordination will be referred as *hosted-T2*. Moreover, as basic case, the relations of T2s with LCG as a whole will be regulated by special agreement with the hosting-T1. This status assumes that the hosting-T1 should coordinate the elaboration of the GRID service requirements by the hosted-T2. This requirement is strengthened also by other requirements, on storage and CPU resources allocated at hosting-T1 for hosted-T2 needs, because these resources should be operated as a part of the GRID.

[3] http://lcg.web.cern.ch/LCG/PEB/MoU/Report_to_the_MoU_Task_Force.doc
[4] http://www.eu-egee.org/
[5] http://www.ivdgl.org/grid2003/

As a result, one should consider both hierarchical structure, T1-T2-T3-…, of the LHC regional centres and the GRID infrastructure (sometimes referred on as *GRID cloud*) when one discusses the GRID services requirements to be provided for T2s.

A number of GOCs are planned to be created around the world (some have started the operation already). At CERN the EGEE *Operations Management Centre* will be created as a part of the EGEE infrastructure.

Then, according to the EGEE plan there will be nine ROCs in Europe, located in each of the national or regional EGEE federations: at CCLRC-RAL (UK), at CC-IN2P3-Lyon (France), distributed ROC in Italy (INFN and some universities), distributed ROC in Sweden (SweGrid centers) and The Netherlands (FOM), distributed ROC in Spain (IFAE and CSIC) and Portugal (LIP), distributed ROC in Germany (FZK and GSI), distributed ROC in South East Europe (in Greece, Israel and Cyprus), at CYFRONET (Poland), and distributed ROC in Russia. Then, five CICs are under creation in Europe: at CERN, CCLRC-RAL (UK), CC-IN2P3-Lyon (France), INFN-CNAF (Italy), and SINP MSU (Russia).

In USA currently the Indiana University is operating as a GOC for GRID3, and distributed model is under discussion for future GOC in OSG (Open Science GRID).

In Asia there is a plan to create GOC, probably in Taiwan.

One should add that the LHC experiments could request GRID services to different T1s or even to T2s.

The services to be provided by GOCs for resource centres, thus to T2s, can be shortly described by referring to the EGEE formulations for ROCs and CICs:

> "the ROCs must assist resources in the transition to GRID participation, through the deployment of GRID middleware and the development of procedures and capabilities to operate those resources as part of the GRID. Once connected to the GRID, the organizations will need further support from the ROCs to resolve operational problems as they arise."

> "Core infrastructure manages the day-to-day operation of the GRID, including the active monitoring of the infrastructure and the resource centers, and takes appropriate action to protect the GRID form the effect of failing components and to recover from operational problems. The primary responsibility of EGEE CICs is to operate essential GRID services, such as databases used for replica metadata catalogues and VO administration, resource brokers (workload managers), information services, resource and usage monitoring."

Thus, the following scenarios can take place for the T2 centre:

- the hosting-T1 is one of the GOCs and provides all necessary services (examples – CCLRC-RAL in UK, and CC-IN2P3-Lyon in France);
- the GRID services are provided by GOCs hosted at different T1s, including the hosting-T1;
- the hosting-T1 has no functions to provide core or operation GRID services for some experiments;
- some of the GRID services are provided by GOC team at the T2 centre itself;

- some of the GRID services are provided by GOCs teams at the T2s which are *brothers* (being hosted by the same T1).

The described above possible scenarios resumes to the following requirements to the hosting-T1:

Hosting-T1, together with T2 hosted, should define the map of GOCs, which will provide necessary GRID services for this T2.

The hosting-T1 should participate in preparing corresponding agreements with defined GOCs, based on current SLAs, if necessary with inclusion of specifics of these particular T2 and T1.

Then, the hosting-T1 should help the hosted-T2 to update these agreements, if it is asked by new GRID releases.

Finally we give the list of basic GRID services to be provided for a T2 centre by the defined GOCs:

1. Distribution, installation and maintenance of GRID middleware and other special systems software and tools. Validation and certification of the installed middleware;
2. Monitoring and publication of the GRID enabled resources (disk/tape, CPU and networking) at hosted T2, including resources allocated at hosting-T1 for the hosted-T2 needs;
3. GRID enabled resources usage accounting;
4. Monitoring of GRID services performance for hosted T2, to ensure the agreed QoS;
5. Provide core GRID services, such as databases for replica metadata catalogues, resource brokers, information services, support of the experiment VOs services, ensure basic GRID security services etc.;
6. Support of GRID specialized networking solutions for effective point-to-point (or few-to-few) connectivity.

5. Support Requirements

As the LCG environment develops, it is recognized that a variety of support functions analogous to the support functions found in computer centre helpdesks, software support organizations, and application development services, will be needed supporting the globally distributed users. The main support functions include well defined **Help Desk Processes, User Information and Tools, Service Level Agreements, User Accounts and Allocation Procedures, Education, Training and Documentation, Support Staff Information and Tools,** and **Measuring Success**. In this chapter is defined, which of these support functions should be provided by the T1 centres and which by the Global Grid User Support Team(s) (GUS).

Help Desk Processes (GUS): As mentioned earlier, every T2 should be hosted by a T1. The hosting T1 has to support operations, sysadmins and users of the hosted T2s, because most T2s are quite small and can't provide all these services themselves. But most of support functions should be provided mainly by the Grid User Support Centres (GUS) and also by the Grid Operations Centres (GOC), of which there will be three each distributed around the globe and which will take on duties for a larger group or even for all T1s and T2s.

The GUSs will provide a single point of contact for all the LCG users, via web, mail or phone. The task of the GUSs will be to collect and to respond as a globally organized user help desk the various user problems. For this the GUSs use a centralised ticketing system. Normally the end users will first call the experiment user support (ESUS), and the experiments are responsible for providing help desk facilities as a first level user support service for their collaborators. The ESUS people will filter out the experiment specific problems and send the remaining problems to the GUS's people. At the GUSs all problems will be written in a problem database, which will also get the concerning solutions; in this way this problem database is becoming a knowledge database containing all known problems. This knowledge database should be accessible not only for T1 and T2 staff but also for all end users. Concerning the experiment software installation and support it is the responsibility of the experiments to ensure that they have sufficient support for this to cover the T0, T1, and T2 centres.

It is agreed that experiment software installation, and its validation, is a responsibility from the experiment, not from the T1. The T1 could act as "contact point" or "link" between the experiment and the lower Tiers/RCs so that the experiments don't have to talk with hundreds of sites.

Following the report of a GDB working group concerning GUS the globally distributed GUS will provide a support for the users on a 24/7 base. It is therefore not necessary that the T1s also provide an around-the-clock availability of their specialists.[6]

User Information and Tools: This is the provision of important information resources and tools to enable the use of the LCG environment and ranges from basic online documentation to information about the current status of resource in the LCG environment and the LCG infrastructure itself to debugging and performance analysis tools. This also includes the methods of delivery of these information resources and tools. This service should be given by the GUSs.

Service Level Agreements: It is important for the LCG providing the grid environment to appropriately set the shared expectations the users of these environments and those providing support. A clear statement that accurately delineates these expectations for both the users and support operations in a grid computing environment is therefore critical and should be elaborated.

User Accounts and Allocation Procedures: All LCG users need to obtain some type of account and some form of authorization to use specific resources within the LCG environment. Accounts should be given by the T1s for their own users and also for the

[6] http://lcg.web.cern.ch/LCG/PEB/gdb/WG5/WG5-Report-V2.0.doc

users of their hosted T2s. The rules for establishing accounts are elaborated by a specific GDB working group.[7]

Education, Training , and Documentation: The LCG users need to be educated and trained in it's use. Ideally, if a user is trained how to work in the LCG environment, this will mean the user will not have to learn the individual nuances of using all of the various resources within the LCG. In practice, this goal may be difficult to achieve, so the need for instruction on some "local" issues for resources on the LCG will likely need to be maintained. Nonetheless, what is new to the majority of users is the distributed grid environment and, just as documentation of this is needed, training is required to develop a user community fluent in the use the environment. This include both on-line and in-person training activities. Nevertheless basic training and tutorials normally should be provided in a centralised manner by CERN. Following this the T1s should provide support and documentation for deployment and maintenance of these grid software packages for "their" T2 people.

Support Staff Information and Tools: The support staff must have at their disposal a number of “tools of the trade” and information resources to effectively provide support to the user community. This include such things as the GUS knowledge base to draw upon, information about the status and scheduling of resources and grid services, tools to assist in the diagnoses of problems reported, and appropriate levels of access to resources to operate effectively. By now it's not yet totally clear, which part of this work has to be done by the GUSs and which one by the GOCs. This service should be given in a shared manner by the GUSs and the T1s.

Measuring Success: The support groups at the Grid User Support and in the T1s need some way to determine success or failure of problem solving and support methods. This is seldom an easy task because it can be largely subjective. While qualitative information is a more useful indicator of the success of the support organization, it is more difficult to get. Frequently, this information can be obtained from various forms of user feedback. One possible way to this could be to collect quantitative metrics, which are fairly easy to collect. Effective measures must be in place to advance the support functions. It seems to be necessary to seek even more effective and accurate indicators of the performance of the GUS and GOC support groups.

[7] http://agenda.cern.ch/askArchive.php?base=agenda&categ=a04113&id=a04113s1t1/document