Available on the CMS information server **CMS NOTE - 2006 / 128**

The Compact Muon Solenoid Experiment

CMS Note

Mailing address: CMS CERN, CH-1211 GENEVA 23, Switzerland

15th June 2006

CPT White Paper on Tier-1 Computing Resource Needs

The CPT Project

Abstract

In the summer of 2005, CMS like the other LHC experiments published a Computing Technical Design Report (C-TDR) for the LHCC, which describes the CMS computing models as a distributed system of Tier-0, Tier-1, and Tier-2 regional computing centers, and the CERN analysis facility, the CMS-CAF. The C-TDR contains information on resource needs for the different computing tiers that are derived from a set of input assumptions and desiderata on how to achieve high-throughput and a robust computing environment.

At the CERN Computing Resources Review Board meeting in October 2005, the funding agencies agreed on a Memorandum of Understanding (MoU) describing the worldwide collaboration on LHC computing (WLCG). In preparation for this meeting the LCG project had put together information from countries regarding their pledges for computing resources at Tier-1 and Tier-2 centers. These pledges include the amount of CPU power, disk storage, tape storage library space, and network connectivity for each of the LHC experiment for the subsequent five years.

In this White Paper we describe the current situation for CMS regarding pledged computing resources.

Executive Summary

CMS plans to use Tier-1 resources foremost to provide its one active, custodial copy of event and MC data for high-throughput access and long-term storage. Within the Worldwide LHC Computing Grid, seven countries contribute Tier-1 resources to CMS in terms of CPU, disk and tape resources. CMS has identified a set or problems related to the sharing of the available resources in the WLCG:

- For CMS there is a 42% shortfall in the required storage space for data samples, meaning that CMS has been unable to identify enough storage resources to store even the single active copy of the total sample of event and MC data.
- There is a distortion in resources offered to different experiments versus the needs specified in the C-TDRs. For example, CMS lacks about half of the (relatively cheap) tape capacity for its custodial data storage, while for Atlas there is an abundance of (more expensive) disk to store duplicates of some data samples.
- There is a large unbalance in resources between Atlas and CMS. The total computing funding and available resources to Atlas and CMS computing within the WLCG is different by factors of two. This is despite the identical missions and very similar operating conditions of the experiments.
- The process of allocating computing resources for the LHC program is ill-defined.

This white paper analyses these issues in more detail.

In summary, it is recognized that both experiments, Atlas and CMS, have identical missions and require approximately equal resources. This point of view has been endorsed and put forward by the LHCC in its reviews of the computing models of the experiments. Given the current pledges from Tier-1 countries to CMS and Atlas the funding for each experiment is quite different.

We also note that the needs for computing resources were determined within stringent constraints, always cutting into physics. In such a resource-limited environment the experiment requests for computing resources should not be taken as truly fundamental requirements.

CMS is proposing to establish a mechanism for balancing and sharing of resources between experiments. This should be a process under the auspices of CERN as host laboratory, in consultation with the Tier-1 agencies. Input into the process would be the funding envelope given to experiments by the funding agencies. Experiments would then work with sites to define the optimal deployment and balance of resources, and work with individual agencies to increase resources in case of shortfalls, especially for storage and tape, for example by making more use of existing tape installations.

A peer review should then evaluate the choices of technical parameters, within the constraints of the well-defined resource envelope. The process should allow for iterations to adjust parameters and potentially increase the funding envelope in order to optimize the overall LHC research program.

1. Assessment of the CMS Computing Resource Situation

1.1 Computing Model and Role of Computing Tiers

The CMS Computing Model is based on a hierarchical set of tiers of regional computing centers, with well-defined roles and functions for the CMS data flows and work flows for tasks such as event processing, calibration running, data access and skimming, data analysis, and MC simulations.

Role of Tier-0 and the CMS-CAF

At CERN, the Tier-0 and the CMS CERN Analysis Facility (CMS-CAF) provide complementary functions for CMS computing, being different “logical” entities within a single large “physical entity”. Only together do they provide the required CMS computing at CERN.

The Tier-0 is for highly organized “production work” with quasi-real time data flows. This includes prompt reconstruction, prompt calibration, re-processing of express data streams, etc. The Tier-0 will not support logins from general users of CMS but only those carrying out specific production related activities.

The CMS-CAF is for high-priority asynchronous access to data coming from the Tier-0, to perform verification of the detector and trigger performance and calibration, for data quality assurance, for rapid analysis of high-priority or “express line” physics (5-10% of all data taken), and for data analysis by users at CERN.

LHC running time is precious, so it is imperative that CMS can do fast and efficient monitoring of the quality of the data taken. Beyond the detector monitoring it is extremely important to monitor the trigger, with ad-hoc studies of detector data (using special data streams) and with a few critical analyses that verify physics, such as looking at mass spectra, cross sections, etc. This is an important function of the CMS-CAF. For calibration and alignment CMS needs to have fast turn-around for both the Tier-0, and potentially the HLT, with short latency and fast turn-around - hours rather than days.

There is also an important role for the CMS-CAF in enabling fast access to the data for physics assurance and analysis. This is to check whether CMS sees indications of something unexpected (either background or signal) that might require immediate action, such as a trigger adjustment, and to do rapid analysis of express-line physics to extract hot physics results. The role of the CMS-CAF is particularly important during the LHC startup phase.

The resource needs for Tier-0 and the CMS-CAF are estimated in the CMS Computing TDR[1]. The Tier-0 resources are highly optimized for prompt data processing, streaming to tape and distribution of data to Tier-1 centers. The CMS-CAF resources comprise an analysis farm with access to a subset of the data produced at the Tier-0. It is about the same size as a nominal Tier-1 plus 2.5 nominal Tier-2 centers. The proximity of the CMS-CAF to the Tier-0 reduces the storage needs with respect to regular Tier-1/Tier-2 centers, as most data are directly accessible from tape and disk at the Tier-0.

Table 1 shows the required ramp up of resources for the Tier-0 and the CMS-CAF over the first years of running in 2008 to 2010.

[1] “CMS Computing Technical Design Report”, The CMS Collaboration, CMS TDR 7 and CERN / LHCC 2005-023, (2005).

		2008	2009	2010
Tier-0	CPU [MSI2k]	4.6	6.9	11.5
	Disk [PB]	0.4	0.4	0.6
	Tape [PB]	4.9	9.0	12.0
CMS-CAF	CPU [MSI2k]	4.8	7.3	12.9
	Disk [PB]	1.5	2.5	3.7
	Tape [PB]	1.9	3.3	4.8

Table 1 Required ramp up of resources for the Tier-0 and the CMS-CAF over the first years of running in 2008 to 2010.

Role of Tier-1 centers

The Tier-1 centers provide dedicated computing facilities to store the single active copy of CMS real and simulated data, for subsequent reprocessing, skimming, event serving and other large-scale tasks that require fast access to the bulk data. CMS Tier-1 resources do not include data analysis resources and MC production resources, as these are the purview of the Tier-2 centers[2].

Real data is transferred to the Tier-1s from the Tier-0 after prompt reconstruction. MC data comes to the Tier-1s from the Tier-2s and Grid CPU resources, where MC is being produced in a distributed fashion. A major responsibility of the Tier-1 centers is to provide permanent storage for this real and simulated data. This should be in a form that allows high-throughput access for providing selected subsets of data to Tier-2 centers, where individual users run analysis jobs.

At the Tier-1 centers CMS requires hierarchical mass storage and data access systems. These use tape library back-ends for custodial data storage with high-throughput disk caches. In many cases full datasets are “pinned” on the front-end disk storage systems. Enough CPU needs to be provided to allow re-processing of data when new calibration and detector alignment information becomes available, and to allow the fast skimming of the hosted simulated and real datasets to extract the relevant samples and send them to Tier-2s for further analysis.

Role of Tier-2 centers

CMS will use a set of Tier-2 centers that are smaller but more numerous than Tier-1s. They have substantial CPU resources to support analysis, calibration activities and Monte Carlo simulation. Tier-2s rely on Tier-1s for their access to large datasets and for secure storage of new data they produce.

The basic functions supported by a Tier-2 include:

- fast and detailed Monte Carlo event generation;
- data processing for physics analysis, with very fast data access for late stage analyses;
- data processing for calibration and alignment tasks, and detector studies.

[2] It is understood that some Tier-1 sites may chose to also provide analysis capabilities which are not accounted in the CMS Tier-1 resource requirements.

In 2008 a nominal Tier-2 center will provide CPU resources of 0.9 MSI2k and 200 TB of disk storage. Tier-2 centers also require excellent networking connections of at least 1 Gbps, to enable data transfers to and from the Tier-2 center and any of the CMS Tier-1 centers.

Data Flows

The main data flows are: the import of data sets such as AOD (Analysis Object Data) from the closest Tier-1 center; the transfer of skimmed RECO or FEVT data from any of the Tier-1 centers that host the corresponding primary datasets; and, at a lower rate, the movement of MC files which were produced at a Tier-2 to a Tier-1 center for storage. For MC datasets the production will typically be distributed and several Tier-2 sites will contribute. Therefore individual data files from Tier-2 centers will be consolidated into a single CMS dataset that is hosted in the tape library of a Tier-1 center chosen by the CMS data management system. From there it will be available for further data analysis to any Tier-2 center.

Datasets for physics analysis will come from the Tier-1 center which provides easiest access in a metric determined by the data management system. If AOD data is requested each Tier-1 center has the full sample, so it will be pulled from the "nearest" Tier-1 center, taking into account network throughputs and load factors at the Tier-1 centers. In case of RECO or FEVT skimmed datasets, which will be a very important use case for the initial years of running, data will be pulled from the Tier-1 center that has that particular dataset (of which there will typically be only one copy available in the whole system of Tier-1 center). From this it follows that the connectivity of Tier-2 centers to *each* Tier-1 center is required.

Given the complexity of current network topologies, multi-hop data transfers involving more than one Tier-1 center may be required to achieve the required throughput of moving data in and out of Tier-2 centers. As long as this functionality is not available on the network level or as part of the data transfer Grid services, such multi-hop transfers can be implemented in the CMS data management system as part of the data placement service.

In many cases Tier-1 and Tier-2 centers will have specific organizational relationships with each other, in the scope of national computing projects or local Grids, through support agreements or other arrangements. Also, as part of the WLCG fabric, Tier-2 centers may negotiate and build up excellent networking connection to a particular Tier-1 center. Such agreements will be very helpful to the WLCG infrastructure and the overall throughput in the CMS grid system. However, these should be negotiated as part of the WLCG collaboration and are outside the scope of the CMS experiment itself.

Figure 1 summarizes the data flows between the different tiers in the CMS computing model, indicating required rates and throughputs.

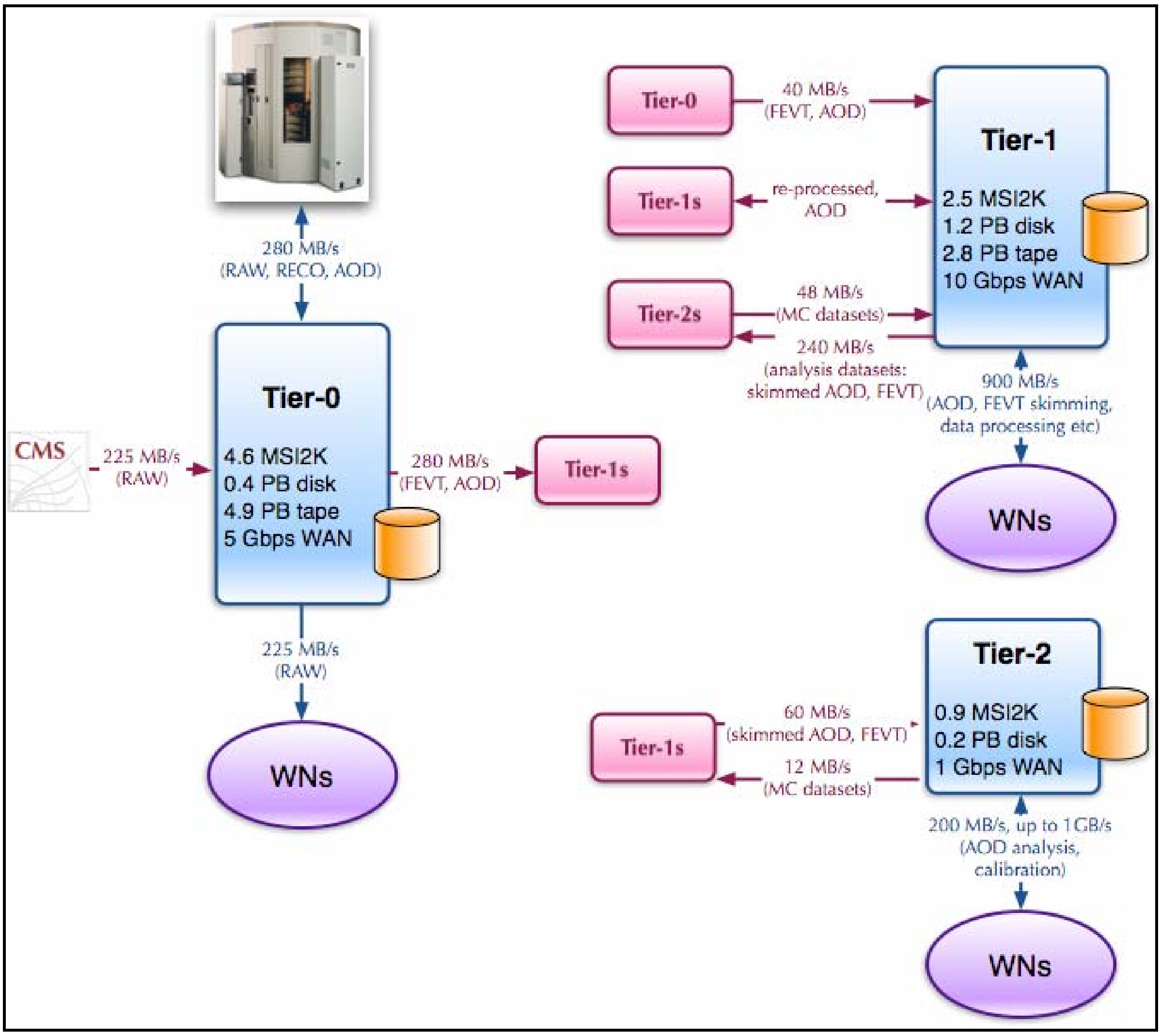

Figure 1 CMS data flows between Tier-0, Tier-1, and Tier-2 centers. Horizontal arrows indicate flows into/out of computing centers, vertical arrows indicate main data flows internal to them (to local CPU's, called WNs) and (shown explicitly only for the Tier-0) to mass storage (tape silo picture). In this picture it is assumed that a Tier-1 manages traffic corresponding to about four Tier-2's.

Default connections of Tier 2 Centers to Tier-1 Centres

Table 2 lists the known CMS Tier-2 centers (as of April 2006) and the proposed default connection to the nearest Tier-1, where it is known.

Country hosting Tier-2(s)	Nearest Tier-1 (if known)
Belgium	IN2P3
Brazil	Fermilab
China	
Croatia	
Estonia	RAL
Finland	
France	IN2P3
Germany	GridKa
Greece	CNAF
Hungary	
India	ASGC
Italy	CNAF
Korea	
Pakistan	ASGC
Poland	GridKa
Portugal	PIC
Russia	
Spain	PIC
Switzerland	GridKa
Taipei	ASGC
UK	RAL
USA	Fermilab

Table 2 Current set of CMS Tier-2 centers (as of April 2006) and the proposed default connection to the nearest Tier-1, where known.

1.2 Requested Resources

Table 3 shows the resources requested by CMS to process and analyze data in 2008, the year of the first full physics run of the LHC, at the CERN Tier-0 and CMS-CAF, the Tier-1 centers and the Tier-2 centers (from Table 5.3 of the C-TDR).

Tier	CPU [MSI2k]	Disk [PB]	Tape [PB]
Tier-0	4.6	0.4	4.9
Tier-1	15.2	7.0	16.7
CMS-CAF	4.8	1.5	1.9
Tier-2	19.3	4.9	---

Table 3 Resources required to process and analyze CMS data in 2008 at the Tier-0, CMS-CAF, the sum of Tier-1s, and the sum of Tier-2 centers.

This note concentrates on Tier-1 resources. There are seven Tier-1 centers that pledge resources to the CMS experiment. These centers are in France, Germany, Italy, Spain, Taipei, the UK, and the USA. Very large resources are available at these centers with a total of about 30 MSI2k CPU, 15 PB of disk and 15 PB of tape space. In centers that do not cater to CMS there are additional 5.5 MSI2k CPU, 3.1 PB of disk and 2.3 PB of tape space available.

Tier-1 Resources and Nominal Tier-1 Sizes

As a guideline CMS has expressed the resource needs at Tier-1 centers in terms of a "nominal" Tier-1 center. The total resource needs are six such nominally sized centers plus the resources requested at CERN. In reality there will be variations from this nominal size (and perhaps the total number of Tier-1s), depending on the magnitude of the resources that individual countries can contribute to CMS.

For the 2008 running, each nominal Tier-1 center will host 0.5 PB of event and 0.5 PB of simulated data samples. Importantly each one also hosts 0.6 PB of analysis data that is to be served to Tier-2s (each Tier-1 keeps the complete sample including re-runs). Also there is 0.5 PB of storage required for re-processed samples that need to be kept accessible for some time in the tape library (re-processing of its share of event and simulated data is a main task of the Tier-1 centers). In addition there needs to be some space for calibration samples and analysis running. The data samples taken in 2007 also need to be held in custodial storage.

In the CMS computing model, tape resources correspond to the safe storage space that Tier-1 centers must provide to the experiment to store a secure custodial copy of the data. This is for the single copy of event and MC data that is accessible for data processing at high throughputs and small latencies[3].

To achieve sufficient throughput for data analysis and processing, CMS requires a certain amount of disk resources that either have copies of data pinned on disk or that function as caching disks as part of the hierarchical mass storage system.

A nominal Tier-1 center provides enough CPU to do the re-processing of event and simulated data and to allow skimming and some high-throughput analysis running on the data samples that it hosts.

Minimal Tier-1 Sizes

There is a minimal size for a Tier-1 center to be functional as part of the CMS computing model. CMS requires a Tier-1 center to have (at least) sufficient resources to host about half of the above sample sizes of real and simulated datasets, and a full copy of the AOD samples of real and simulated data (only one rerun), together with at least one re-processed sample of these data samples. The CPU requirement of the minimal Tier-1 center is correspondingly scaled to perform the required processing of these smaller datasets.

Table 4 shows the resulting numbers, allowing for efficiency factors, for 2008 for a single nominal Tier-1 (assuming 6 in total) and the minimum capacity for a Tier-1 to be functional.

Size	**CPU [MSI2k]**	**Disk [PB]**	**Tape [PB]**
nominal	2.5	1.2	2.8
minimal	1.3	0.7	1.2

Table 4 Nominal and minimal sizes of a CMS Tier-1 center in 2008, assuming a total of 6 Tier-1 centers.

[3] For event data there is also a second *inactive* copy at the CERN Tier-0, which is for backup purposes and re-processing after the end of the running period. It is not available for data processing due to the lack of sufficient resources at CERN.

Table 5 shows the required ramp-up of the nominal Tier-1 center capacities from 2008 to 2010.

Year	CPU [MSI2k]	Disk [PB]	Tape [PB]
2008	2.5	1.2	2.8
2009	3.5	1.7	4.9
2010	6.8	2.6	7.0

Table 5 Ramp-up, from 2008 to 2010, of a single nominal Tier-1 center capacity, assuming a total of 6 Tier-1 centers.

Tier-1 Resources Pledged to CMS

For 2008, the following total amount of Tier-1 resources are pledged to the CMS experiment:

- 11.6 MSI2k CPU
- 5.5 PB disk, and
- 9.6 PB of tape storage.

Table 6 shows the resource pledges at each of the Tier-1 centers for the data taking year 2008. The percentages denote the resource pledge compared to the nominal Tier-1 size for that year.

Tier-1 Country	CPU (% nominal)	Disk (% nominal)	Tape (% nominal)
France	1.5 MSI2k 60%	0.8 PB 65%	1.2 PB 42%
Germany	1.2 MSI2k 48%	0.7 PB 54%	0.9 PB 32%
Italy	1.9 MSI2k 77%	0.9 PB 73%	0.7 PB 26%
Spain	0.8 MSI2k 30%	0.4 PB 29%	0.8 PB 30%
Taipei	1.5 MSI2k 61%	0.7 PB 56%	0.6 PB 21%
UK	0.4 MSI2k 18%	0.2 PB 19%	0.7 PB 24%
US	4.3 MSI2k 170%	2.0 PB 166%	4.7 PB 168%

Table 6 Resource pledges at each of the Tier-1 centers for the data taking year of 2008.; the percentages denote the pledge compared to the nominal Tier-1 size.

Despite the large resources available overall at Tier-1 centers, there is a shortfall for CMS of a little more than 20% for both CPU and disk, and a shortfall of about 40% (8.5 Petabyte missing) in the required tape capacity.

Table 6 shows that:

- only one single Tier-1 Center (in the USA) is at least the size of a nominal Tier-1 center;
- the smaller centers reach just 20-30% of the required nominal CPU and disk storage size;
- the pledged tape resources are disproportionately small, around 30% of the nominal size or below even for the larger centers (except for the USA and France).

For following table shows the estimated resources available to CMS at the Tier-1 centers for the years of 2008 to 2010:

Year		**2008**	**2009**	**2010**
CPU (MSi2k)	est. pledges	11.6	19.3	29.5
	requested	15.2	20.7	40.7
	balance	**-24%**	**-7%**	**-28%**
Disk (PB)	est. pledges	5.5	8.8	12.6
	requested	7.0	10.5	15.7
	balance	**-21%**	**-16%**	**-20%**
Tape (PB)	est. pledges	9.6	17.3	24.9
	requested	16.7	29.5	42.3
	balance	**-42%**	**-41%**	**-41%**

Table 7 Estimated total Tier-1 resources pledged for 2008-2010, compared to requirements.

1.3 Analysis of Tier-1 Pledges

CMS has expressed its processing and resource needs in terms of a “nominal Tier-1 center”. It is a big problem for CMS computing that only one of the seven centers actually reaches nominal size (the U.S. center). All the other centers are typically very much smaller than nominal size, down to 20% of a nominal center. In the case of the UK, the CPU pledged does not even reach nominal Tier-2 size, the disk barely reaches the nominal Tier-2 size, and the tape is only sufficient to store the AOD data.

With respect to the core function of the Tier-1, namely as an ensemble of centers to provide custodial storage and access for the *only active copy of the event data*, the centers fall dramatically short of required needs. Only half of the required tape storage is pledged. All centers provide only about 30% or less of the nominal tape storage size, except France (at 42% of nominal) and the U.S. (at 170% of nominal). In total there are 8.5 PB of custodial tape storage space missing.

To put this lack of data storage capacity into perspective we show in the table of required data sample sizes at a nominal Tier-1:

Data Class	**Sample Size at Nominal Tier-1**
Event Data FEVT	0.5 PB
Simulated Data	0.5 PB
Re-processed Data RECO	0.5 PB
Full AOD Samples	0.6 PB

Table 8 Data samples sizes in 2008 for a nominal Tier-1.

It is thus a major concern that 5 out of 7 CMS Tier-1 centers pledge tape libraries that are significantly smaller than (and sometimes only half of) 1 PB in size. Except for the USA, each center provides tiny capacity fractions of just between 4% and 7%. These sizes are enough to just store the AOD data to serve them to Tier-2 centers, but certainly not enough to fulfill the main function, to provide the active copy of the CMS event data and the MC data.

In terms of CPU and disk sizes, 4 out of 7 centers provide less than 10% of the total resources each (the smallest being a mere 3%), and thus are less than or about half the nominal Tier-1 size.

Uneven Distribution of Resources

CMS has analyzed how this uneven distribution of resource pledges between centers and between experiments comes about.

Firstly: it should be noted that only those countries that plan on running a Tier-1 center contribute to the Tier-1 resources; therefore the majority of the 40 CMS countries pay nothing towards the Tier-1s.

Secondly: although ten countries provide Tier-1 centers as part of the WLCG, only seven countries have pledged Tier-1 resources to CMS, namely those where CMS has collaborators.

Thirdly: regional decisions on how resources are distributed between LHC experiments are different in different countries. At the two extremes between the different models employed there are:

- the model of up-front equal funding amounts, ignoring the (large) difference in the relative numbers of authors in that country (e.g. as for ATLAS and CMS in the U.S.);
- the pro-rata scaling of a total amount or resources funded at a given center, according to (1) the fraction of participation in the experiment in that country and (2) the requested resources in the respective C-TDRs (as used in Europe).

These effects can leads to big imbalances in the resources pledged to different experiments. Unfortunately both the above effects disfavor CMS with respect to other experiments.

The first unfavorable factor for CMS is its demographics. CMS has a 40% population from countries that do not provide any Tier-1 resources. In addition it has a relatively large (30%) fraction of collaborators from the USA but does derive a corresponding pro-rata benefit due to the fixed equality of funding between ATLAS and CMS.

Only 30% of the CMS collaboration is located in the other (non-US) Tier-1 countries, which is small compared to the other experiments. Since these countries typically share resources on a pro-rata basis, CMS receives correspondingly less.

In the current scheme of the WLCG, Tier-1 contributions from countries are not treated as a “common fund” - only those countries that have a Tier-1 center contribute. Given the rather different demographics between Atlas and CMS, Atlas enjoys resources from countries that do not at all contribute to CMS. This fraction is very sizable, in terms of CPU and disk space it corresponds to about 50% of the total pledge to CMS.

As described in the next section, the other factor, in addition to demographics, is that the estimated resource needs quoted in the C-TDRs for CMS and ATLAS were not derived on an equal basis of assumptions. They are should not therefore have been used for scaling the experiment contributions in any given country. The different estimates are not directly comparable due to the striking differences in the input assumptions and thus the derived resource needs.

Process for Resource Pledges to the WLCG

We maintain that there was a disconnect between the process to determine resource pledges to the WLCG and the requests of the experiments. During the summer of 2005, when these pledges were determined there was no process within experiments to solicit pledges based on the experiment requests. Moreover, there was no scrutiny of the relative significance of C-TDR numbers that were used at face value to determine resource ratios at the WLCG Tier-1 centers. Instead the LCG project book-keeping process, driven by the “Phase-2 Planning” group, directly went into tables that then were presented to the C-RRB.

It is our view that the process to arrive at these resource pledges is fundamentally flawed. The process leading to the RRB pledges had no mechanism to arrive at a fair resource sharing between experiments. Although countries have used the (very differently sized) claimed resource needs as input into their decisions on allocating resources between experiments, these inputs (and the assumptions underlying them) were not reviewed and put into perspective with respect to each other.

Also, neither has there been a funding or resource envelope guidance to the experiments that would have allowed them to optimize choices of parameters in the model and to come to comparable resource needs, nor has there been any guidance on the choice of (physics or technology driven) input parameters in the computing models. Both could have led to comparable compromises in physics performance between the experiments. The lack of such common guidance means that the needs of the experiments, as quoted in the C-TDRs, are not directly comparable.

The process of arriving at these pledges from the funding agencies has been done during the first half of 2005 *not* within the experiments, but as part of an administrative process within the LCG project. Regional centers have been solicited for resource pledges as part of the "LCG Phase 2 Planning" meetings. The tables including the resource distributions between experiments have been distributed as a "fait accompli", without any review process. This happened during the summer of 2005 even before the resource requests stated in the C-TDRs of the experiments had been reviewed and put into perspective.

As a matter of fact, the first cursory report from the LHCC review of the C-TDRs was only presented in the same computing RRB meeting that agreed on the MoU.

The preliminary observations of the LHCC shown in that meeting stated that

"... the LHCC does not see any significant difference in the fundamental computing needs between ATLAS and CMS. The differences in their requests depend mainly on the details of the computing model in the use of disk versus tape. This is a concern and must be resolved".

However, even now the big imbalance between different experiments is still unresolved. The total funding going to each of Atlas and CMS is of very differing size - both the pledges and resource needs for Atlas are almost a factor of two larger than those of CMS.

There is also a large distortion between the pledged resources *vs*. the specified needs. For example, CMS lacks about half of the tape storage needed for the *only* custodial storage of CMS real data at Tier-1 sites. At the same time there is an apparent abundance of (more expensive) disk space at the same centers to support multiple copies of event summary data samples for Atlas. Finally, it does not appear that the "physics output per additional CHF investment" is the same between experiments that are fundamentally rather comparable in their respective physics reach.

2. Computing Model Baseline Choices: Effect on Resources

The CMS computing model provides a baseline of input parameters and choices. These choices are based on assumptions the validity of which in many cases is described in the CMS C-TDR. These assumptions then result in the CMS resource needs in terms of CPU, disk storage, tape storage that are reported in the C-TDR.

2.1 Assumptions on input parameters

Input parameters to the model include, for example: trigger and data rates, event sizes for RAW, RECO, AOD, and CPU needed to process or simulate an event. These assumptions are rendered as single numbers, but in most cases are chosen from within a spectrum of possibilities, where the exact forms of these distributions (e.g. the RAW event size as function of luminosity etc) are our current best estimates. Therefore, these parameters generally have significant uncertainties.

Importantly, at the LHC the choices of these parameters should be seen as “cuts” that seek to maximize the physics opportunity through judicious compromises that reflect resource realities. The CMS baseline choices for these parameters are, in many cases, more restrictive and cut deeper into the physics than the choices of Atlas; examples include the trigger rate and event sizes.

2.2 Assumptions on data and workflows

Such assumptions are related to: the need for re-processing; the feasibility of partitioning the data into smaller parts; the definitions of primary datasets; and the ability to actually get the distributed computing model to work in practice on top of a loose collection of computing centers and resources providers.

CMS baseline choices constitute a rather structured data model with "rigid" partitioning of:

- the data into O(50) primary data sets based on (“vertical”) trigger paths, which makes for flexible prioritizing but requires discipline in the choices of triggers and analysis samples; and
- the computing tiers, with defined roles for the Tier-0, 1, 2 centers that allows us to full exploit the resources at these centers, at the price of needing to orchestrate well the workflows across the whole distributed system in a rather hierarchical way.

It is worth noting that there is still flexibility between sites within the same tier and prioritization amongst sites, given the well structured data model, and some flexibility for the higher tiers to contribute to the workload of lower tiers.

2.3 Assumptions on technical parameters

Assumptions such as the "tape-to-disk-ratio", the number of copies of data at the centers, the relative size and contents of the data tiers (e.g. size of the AOD vs RECO), and so on have a significant impact on resource requirements.

The CMS baseline choices are again relatively minimalist. For example, it is foreseen to keep just a single active copy of the FEVT data (although the structured data model allows optimizations like replication of "hot" primary datasets across several sites) and of making flexible use of a hierarchical storage model, where (relatively cheaper) tertiary storage is fully integrated in the computing fabric. This allows flexibility in the use of (relatively expensive) disk storage as a cache, where the cache fraction is a parameter in the system (as opposed to the less flexible "all data on disk" model).

Naturally, in many cases the choices with the best physics yield, the quickest turnaround for physics results and the highest involvement of CMS physicist in data analysis would result just in larger resource needs. The “best” choice for most of these parameters would be “larger is better”, yielding more physics opportunities and ease of operating the computing system.

CMS has generally chosen rather conservative parameter values (in terms of resource needs), being sensitive to realistic limitations on available resources, in particular at the major centers for CMS: the CERN Tier-0 and the Italian and US Tier-1 centers (all other Tier-1 centers are a factor of ~2 or more smaller than these). Still, the resulting resource needs are a factor of 2-4 larger than anticipated in the 2002 CERN computing review (Hoffmann Review). As it turns out, pledges from funding agencies undershoot these requests, in case of data storage needs by a factor of two.

3. Proposed Process for Defining LHC Computing Resources

A possible process for defining LHC computing resources is:

1) A clear statement on the overall funding available for CMS resources at each Tier-1 center and Tier-0 center, expressed as fraction of a "nominal pp experiment" instead of the reported resource needs.

2) A process of peer review on the choice of technical parameters given the resource envelope.

3) A set of requests to sites in terms of how resources are to be provided, with clear guidelines on the level of service and best practices (instead of just specifying technical interfaces and parameters).

4) Iteration of the above steps, adjusting input parameters and technical parameters, also allowing for adjusting the funding envelope for computing

4. Summary

In summary, it is clear that both experiments, Atlas and CMS, have identical mission, so both must require about equal resources. This point of view has been endorsed and put forward by the LHCC in its reviews of the computing models of the experiments. Given the current pledges from Tier-1 countries to CMS and Atlas the funding for each experiment is quite different.

We also note that the needs for computing resource were determined within stringent constraints, always cutting into physics. In such a resource limited environment the experiment requests for computing resources should not be taken as requirements.

CMS is proposing to setup a mechanism for balancing and sharing of resources between experiments. This should be a process under the auspices of host lab, in consultation with Tier-1 agencies. Input into the process will be the funding envelope, given to experiments by the funding agencies. Experiments would then work with sites to define the optimal deployment and balance of resources, and work with individual agencies to increase resources in case of shortfalls, especially for storage and tape, e.g. making use of existing tape installations.

A peer review should be run on choices of technical parameters, given the resource envelope. The process should allow iterations to adjust parameters and eventually increase funding envelope by prioritizing the LHC overall program.