CERN-RRB-2008-106

Computing Resources Scrutiny Group

C.Bozzi (Italy), C.Diaconu (France), D.Espriu (Spain, *Chairman*), J.Flynn (UK), D.Groep (The Netherlands), J.Knobloch (CERN), A.Lazzarini (USA), H.Marten (Germany), W.Trischuk (Canada), B.Vinter (Nordic Grid), H.Renshall (CERN/IT, *Scientific Secretary*)

This report summarizes the deliberations of the Computing Resources Scrutiny Group (CRSG) established by the WLCG Memorandum of Understanding regarding the computing requests by the four LHC experiments for 2009. The purpose of the CRSG is to inform the decisions of the Computing Resources Review Board (C-RRB) for the LHC experiments. The starting point is the resource request information presented to the C-RRB by the different experiments and the guidance that the C-RRB cares to give. The CRSG then enters into a sustained dialogue with each experiment seeking to understand to what extent the computing resource requests are well motivated, the usage made of these resources and the accounting figures regarding usage and availability of the pledged resources.

According to the WLCG MoU, every year the CRSG shall scrutinize

- The resource accounting figures for the preceding year
- The use the experiments made of these resources
- The overall request for resources for every experiment for the following year and forecasts for the subsequent two years
- The CRSG shall also examine the match between the refereed requests and the pledges from the institutions
- The CRSG shall make recommendations concerning apparent under-fundings.

The CRSG held a total of six plenary meetings, sometimes extending over more than one session. It also held several phone conferences and regular contact by email was sustained among all members. A sharepoint web site was established as a document repository. The CRSG contacted the different experiment spokespersons who designated each one or two persons from their respective computing management teams with whom our referees interacted. Two, sometimes three, referees were appointed for each experiment. The CRSG wishes to thank the four experiments ALICE, ATLAS, CMS and LHCb, and in particular their respective computing managers, for the collaboration offered and their remarkable openness.

In carrying out the present scrutiny the scope of this group is largely limited to the implementation of the respective computing models whose TDRs have been reviewed by the LHCC. There is however a gray zone where the respective competences of the LHCC and the CRSG overlap. Furthermore the natural evolution of the commissioning of the experiments as well as the implementation of the computing models in successive tests along with a better understanding of their needs have motivated a number of changes, sometimes representing limitations in the original model or assumptions. When we feel we are not competent to judge the validity or convenience of these changes on the physics side we bring them to the attention of the LHCC.

This scrutiny has been mostly limited to the resources requested for 2009. In order to fulfil the mandate of the C-RRB in this startup period 2008 has been scrutinized too but the exercise in this case is to some extent academic as the resources for 2008 should already be in place by the time this scrutiny is made available to the C-RRB. 2010 and beyond have been partly studied but there are too many unknowns at present to go, at this point, beyond the following general statement: With perhaps a few exceptions, we

have found no gross discrepancies between the resources requested for 2008 and 2009 and those that we believe should be warranted in these years. Based on that, and with the information available at present, we believe that the present extrapolation to future years should largely remain valid. However, for a proper scrutiny we have yet to see real collisions and real data with the computing models going through a reality check. The CRSG prefers not to commit itself to any specific forecast for 2010 and beyond.

The CRSG proposed a standard set of assumptions on beam time. These assumptions have been used for scrutinizing all experiments. They differ considerably from previous scenarios in the case of 2008, and only slightly in 2009. An 'efficiency' of 50% has been assumed in order to extract useful beam time from the total amount that the accelerator will be running. This is an optimistic assumption (recent public presentations suggest that 40% is closer to reality and this is perhaps still too optimistic for the first months of running). The scrutiny used these values

Year	pp	AA
	Beam time (seconds/year)	Beam time (seconds/year)
2008	0.3×10^7	0
2009	0.9×10^7	10^6
2010	10^7	10^6

Some differences between the experiments' requests and our scrutiny arise from the different running conditions assumed.

These beam times would correspond to 3 months of data-taking in 2008 and 7 months of data-taking in 2009 for proton-proton (pp) operations, and 0 months in 2008 and 1 month in 2009 for heavy ion (AA) operations. These were rather optimistic, but attainable, expectations.

Unfortunately, once the scrutiny was completed, the incident in sector 3-4 and the subsequent decision to postpone the restart of the LHC until spring 2009 brought about a reduction to zero of the beam time in 2008. The scrutiny for 2008 has therefore to be understood as an 'exercise' to test the extent of the understanding by the CRSG of the different computing models. Armed with this knowledge we have analyzed in detail the impact of the lack of physics data in 2008 for the 2009 requirements. This is provided as a note added to the different scrutinies (except for ATLAS where the modifications were directly incorporated in the main text). There may actually be an enlarged data-taking period in 2009 as compared to expectations before the incident in sector 3-4 but we believe this extended period is still within the above assumptions for 2009.

Experience gained once real data-taking is underway should reduce remaining uncertainties considerably allowing better estimates for 2010 and beyond. The group also plans to look at the quality and effectiveness of the monitoring and accounting tools in the immediate future.

General recommendations

- It seems prudent to scrutinise the experiments' use of resources after a few months of data taking in 2009. It is also important, given the resource acquisition cycle, to inform the Tier1 and Tier2 computing centres of the resource acquisition plans for calendar year 2010 as soon as possible. The CRSG commits itself to provide a scrutiny at the earliest feasible date and would recommend an earlier CRRB meeting. While it may be difficult in this startup period to suggest definite dates and a substantial advancement may not be feasible in 2009, we think that in future years it would be very helpful to the

funding agencies and the different institutes to have a scrutiny ready by the end of summer, thus giving more time to the Tier1 and Tier2 to complete the procurement process.

- The WLCG represents a computing effort of an unprecedented scale. In spite of increasingly demanding tests being passed uncertainties remain. We recommend that the different collaborations undertake a proper risk analysis and take stock of their results in future requests in order to cope with the most likely failures or shortfalls. We feel that this assessment is particularly worthwhile for two experiments: ALICE and ATLAS for different reasons. In the first case it seems quite difficult for their computing demands to be met and the implications of the under-funding should be understood. In the second case the sheer size of the collaboration and the relatively less organized nature of their computing model makes it more vulnerable.

- In the case of ATLAS and CMS the information provided to us about their AA program has been rather sketchy. While this may not be the main physics goal they are pursuing, and it will impact their 2009 needs in a very limited manner, it will surely have an impact on their future computing needs. We would be thankful to them for more detailed information in successive scrutinies.

- As running conditions may vary in the future (with the presence of 75ns bunch crossings leading to pile-up) the collaborations should be aware that this has to be accommodated within the existing envelope by decreasing the event rate or similar measures.

- The experiments are asked to actively pursue the policy of reducing the size of their raw events, and other derived formats, in future years as much as possible as detectors become better understood.

- A strict policy of removing all ‘dark’ or ‘orphaned’ data should be enforced by the collaborations.

- The CRSG recommends to the experiments to keep their computing models and needs under constant revision. We have found in this first scrutiny a conservative approach according to which some requests had not been officially modified even if it was clear that they were not realistic anymore.

- We recommend the experiments make maximal use of the distributed resources in the GRID avoiding as much as possible the use of CERN facilities.

- In the case of CERN resources, we advocate for a very clear separation between the contributions used for calibration and first pass reconstruction and central analysis (‘express stream’ or similar), and those used to perform physics analysis by the CERN based physicists.

- The CRSG wishes to state that the recommendations contained in this scrutiny are to the best of our knowledge rigorous. They correspond to the real needs of the experiments for a given LHC live time in the present stage of the commissioning and of their computing model implementation. Shortfalls of any kind would seriously jeopardize the success of the experiments. We therefore recommend that the funding agencies ensure the effective and timely delivery of the pledged resources.

LHCC matters

Our scrutiny has identified several aspects that need to be brought to the attention of the LHCC.

- Most experiments propose using increased trigger rates as compared to the ones stated in the TDR reviewed by the LHCC. We feel we are not competent to review the need or convenience to do so.
- ALICE wants to increase substantially their amount of pp data; in particular they stress the benefit of acquiring data at 10TeV. We have not assessed these needs from the physics point of view and we do not know whether such lower energies will be available in the 2009 run or anytime in the future.
- One of our conclusions is to recommend that ALICE undertakes a full assessment of how their physics reach might be affected by requested computing resources not materializing.
- The event size has a very direct impact on the computing requirements. CMS has made an effort to reduce the raw event size (and the size of all subsequent derived formats) by establishing a reduction profile after startup. We believe that this effort should be followed by the experiments with the largest computing needs without unduly jeopardizing the physics.
- We would like to inform the LHCC of potential modifications of the computing models due to the proliferation of different data formats serving the same purposes.
- The realization of the computing model for ATLAS seems to be diverging from the implementation originally envisaged in the TDR for reasons discussed in this report. This implies, in particular, heavier demands on CERN resources. This is surely a matter for the LHCC to examine.
- Cosmic data taking is now much emphasized by experiments; while it is clear that cosmics are extremely useful in commissioning for calibration, this data is by nature transient and it seems somewhat questionable to us to support substantial requests based on cosmic runs, but we do feel we have not sufficient insight to make a definite scientific judgement on this.

Scrutiny of the ALICE Experiment Request

Overview

ALICE aims to establish the existence of and analyse QCD bulk matter and the quark-gluon plasma (QGP). The strategy is to study specific signals for QGP formation or new physics as well as global event information. Comparison will be made between pp and nucleus-nucleus collisions, but there is also an independent pp physics programme relying on ALICE's unique particle ID and low momentum tracking. The detector concentrates on mid-rapidity events with minimum baryon number density and maximum energy density.

The need for comprehensive analysis of nucleus-nucleus collisions (denoted AA or sometimes HI below) with huge numbers of particle tracks coupled with the need for pp analysis makes ALICE's computing requirements very demanding. We have attempted to understand the assumptions and implementation of the ALICE computing model in order to assess the experiment's resources request.

This section summarizes the outcome of discussions with representatives of the ALICE computing management, Federico Carminati and Yves Schutz. We are very grateful to Yves Schutz in particular for patiently answering our questions and for providing us with a copy of the detailed ALICE computing model spreadsheet.

We tried to understand the computing model well enough to verify its major outputs and then compared the ALICE requests with our estimates given the CRSG scenario for LHC operation. Our comments and recommendations follow.

ALICE requests

The ALICE requests for 2008 and 2009 are summarised here. They have not changed since September 2007. In the table “CPU” is given as installed capacity, “MS” denotes custodial mass storage and “Disk” indicates transient storage, while “ext” denotes resources external to CERN. There are T0, T1 and T2 (CAF) resources at CERN.

	Year	CERN	T1 ext	T2 ext	Total
CPU/MSI2k	2008	1.9	10.1	12.5	24.5
	2009	9.7	19.9	14.3	43.8
Disk/PB	2008	1.8	3.9	1.7	7.4
	2009	4.4	6.8	4.0	15.3
MS/PB	2008	3.4	5.7	0	9.1
	2009	7.4	12.4	0	19.7

CRSG commentary and recommendations

The following tables show our estimates for the ALICE needs for 2008 and 2009 together with the experiment’s requests. For storage we checked the ramp-up of requirements. For CPU capacity we checked the steady-state requirement but did not make a detailed check for the ramp-up years.

2008

Resource		CERN	T1 ext	T2 ext	Total
CPU/MSI2k	Request	1.9	10.1	12.5	24.5
Disk/PB	Request	1.8	3.9	1.7	7.4
	CRSG	1.3	4.2	8.9	14.4
MS/PB	request	3.4	5.7	0	9.1
	CRSG	2.8	3.9	0	6.7

2009

Resource		CERN	T1 ext	T2 ext	Total
CPU/MSI2k	request	9.7	19.9	14.3	43.8
Disk/PB	request	4.4	6.8	4.0	15.3
	CRSG	2.5	9.9	9.6	22.1
MS/PB	request	7.4	12.4	0	19.7
	CRSG	7.7	10.6	0	18.3

The raw data volumes for pp and AA running are comparable in the steady-state, though obviously not in 2008. As might be expected, the requirements for AA processing and storage dominate for real data reconstruction, reconstructed data storage (ESD and AOD) and for Monte Carlo (MC) simulation and reconstruction. If handling real data has the highest priority, then varying the amount of MC tasks

provides a means to react to resource shortfalls (or surpluses) during the ramp-up period of LHC operation (but reducing the amount of MC will risk damaging the physics programme).

The ALICE computing model has been well tested for MC simulation and reconstruction. Scheduled analysis using “trains” of analysis tasks has been tested for a month or more. The end-user (chaotic) analysis has been fully exercised (though the number of regular users will grow from its current level of around 60). With the assumptions of the computing model, chaotic analysis makes the least demands on processing power, but is most challenging for data-storage, cataloguing and access strategies.

During the scrutiny period, ALICE implemented zero-suppression for their time projection chamber (TPC) whose output dominates the raw event sizes. Thus the raw data size per particle track is as anticipated in the technical design report (TDR) and pp and AA event sizes will be as anticipated within the assumptions of the event generator for pp and particle multiplicity for AA. For AA the total storage and processing needs are essentially constant (for fixed beam time), being fixed by the bandwidth to Tier 0 mass storage (the number of events recorded is inversely proportional to the multiplicity). For pp running there is flexibility to increase the event rate substantially in the initial running period in 2008 (compared to the steady-state rate) in order to maximise the data taken at lower centre-of-mass energies.

The ALICE requests for 2008 and 2009 look reasonable overall. As shown above we think the disk requirement at T2s is underestimated. In contrast, the mass storage request looks overestimated for 2008, at least partly because the ALICE model is accumulating data from an assumed 2007 startup. For CPU usage, the capacities requested are close to steady-state values in 2009 for CERN and external T1s, with external T2 capacity still ramping up. This seems reasonable given the CRSG assumptions for pp and AA beam time in 2009. In 2008, the CERN CPU capacity is significantly reduced as is appropriate if no first-pass AA reconstruction is needed. The real AA data recorded at the end of 2009 running will be reconstructed during the subsequent shutdown and are available for analysis only thereafter: the current ALICE assumes full AA analysis in 2009. This likely leads to an overestimate of CPU requirement for AA, although AA MC generation and analysis will be done and some analysis of real AA data can be started as soon as sufficient reconstruction has been done during the shutdown.

Experience gained once real data-taking is underway should reduce the remaining uncertainties considerably allowing better estimates for 2010 and beyond.

Our summary comments and recommendations are as follows:

- We assumed ALICE will collect 40% of a standard data-taking year's worth of pp events in 2008 (or from startup) using 30% of a standard year's beam time, allowing an average 33% increase above the long-term pp event rate of 100Hz. ALICE told us that the pp trigger rate could vary from 100 Hz (lower limit to assess detector performance) to around 800 Hz (saturating the bandwidth to mass storage), implying that they could record between 3×10^8 and 2.4×10^9 pp events in 2008. The experiment's intention is to use the maximum event rate the detectors and DAQ will allow.

  There may be good physics reasons to maximise the number of events recorded at 10 TeV, but we believe that the justification to run with an increased event rate at startup has not been reviewed by the LHCC.

- The implementation of zero-suppression for the ALICE TPC during the scrutiny period meant that raw data sizes, or more exactly raw data size per track, are better-known. However, the derived data sizes (ESD and AOD) appear to us to be aspirations (they are input as fractions of the raw event sizes in the

resources calculations). We recommend that these be checked in the light of experience. Since the data volume of AA ESDs is large this is particularly important once AA running commences.

- We verified the steady-state CPU requirements. We were not able to check in detail the ramp-up of CPU requirements but the requests look reasonable.
- Storage requirements were easier to check. We were able to reproduce the experiment's requests, including ramp-up, and consider variations using CRSG assumptions on the inputs. We think the T2 disk storage request is too low and recommend that this be reconsidered by the experiment.
- The combined T0 and T1 mass storage request for 2008 looks generous: it is 36% above our estimate. This appears to be partly because the ALICE spreadsheet accumulates data from an assumed 2007 startup. We appreciate that the computing model as implemented in the spreadsheet was created before the startup date was known, but think it would be wise to rebuild the model using the actual startup date (or best estimate of it) so that the basis for requested resources is more transparent.
- AA data will not be recorded until 2009, although AA MC generation will be running from 2008. The ALICE model assumes full AA analysis in 2009. This is a generous assumption, since although the MC data can be analysed, the real AA data will not start being reconstructed until the winter shutdown and analysis of this only makes sense once enough has been reconstructed to offer meaningful statistics.
- The ALICE model distributes fractions of raw and reconstructed data to T1 and T2 disk storage, with some duplication of reconstructed data (the computing model sends jobs to data, allowing duplication of data which is in high demand). Experience with early running should allow the assumed fractions to be checked and perhaps revised.
- MC production is assumed to be in a 1:1 ratio with real data. AA MC simulation is very demanding and ALICE addresses this by generating underlying AA events which are merged several (ten) times with a signal. Reducing MC production can produce savings in computing resources, but risks compromising physics. It was not clear to us how much reduction in MC production can be tolerated.
- It is by now clear that ALICE's computing requirements are unlikely to be met in practice. We recommend that the experiment make a clear statement to the LHCC how their physics programme will be affected and what can be done to mitigate the consequences of shortfalls.

Note added in response to the absence of 2008 running

We attempted to estimate the effect on 2009 (defined as March 2009 to February 2010) requirements of the loss of 2008 running, using the CRSG assumptions of 9×10^6 and 10^6 seconds of pp and AA beam time respectively. We ignored any requirements for cosmic ray data collection and processing as we assume that this can be safely done with the resources in place in 2008.

2009 will look like a standard data-taking year as far as collecting pp and AA events is concerned. ALICE can easily collect a standard year's worth of pp data from 90% of a standard year of beam time, while the AA beam time is the standard amount.

To make estimates we assumed scheduled analysis of the reconstructed pp data could start after 2 months of running (May) and be complete after 8 months (December),

allowing scheduled analysis of the second reconstruction of the pp data and of the first reconstruction of AA data to start from January 2010 with completion after 6 months. We made a similar assumption for chaotic analysis. We assumed that pp and AA MC generation and reconstruction would carry on throughout the year. This gives a total T1+T2 computing load of half that for a standard year.

CPU Capacity

- The T0capacity will be as for a standard data-taking year, 10 MSI2k.
- If the required T0 capacity is in place from March 2009, there will be significant T1 capacity available at CERN for the period before AA 1st pass reconstruction begins in November 2009 (unused T0 capacity is relabelled and reused as T1 in the ALICE model).
- Assuming analysis starts as soon as sufficient reconstruction has been done, analysis demands should grow during pp running and grow substantially at the end of the period when real AA analysis starts. In particular there would be a high demand for T1 capacity for scheduled analysis at the end of the year.
- With scheduled analysis confined to T1 and chaotic analysis plus MC simulation confined to T2, we estimate that the total CPU capacity over all tiers would be approaching that for a standard data-taking year: we found 43 MSI2k split as 10, 12, 21 between CERN, T1ext, T2ext where "ext" means resources external to CERN (we estimate 51 MSI2k for a standard year). This total happens to match the pre-LHC-incident total request for 2009, although the distribution is somewhat different. Our toy analysis suggests considerable total capacity savings can be made by moving tasks between T1 and T2. By reassigning MC tasks from T2 to T1, for example, we could reduce the total capacity by about 1/3. We commend the experiment to make CPU requirements uniform in time as far as possible.
- Assuming that 40% of a standard data-taking year's worth of pp events had been collected in 2008, and allocating a 3rd reconstruction pass plus scheduled and chaotic analysis of this data before the start of analysis of 2009 AA data, we found an increase in the total T1+T2 computing load to 60% of that of a standard year. However, without attempting to optimise CPU usage over time, we found no effect on the capacity required (it being determined by the AA requirements at the end of the year).
- Initial reconstruction takes place at T0 and we assumed the second reconstruction of pp data would start at T1s in November. This likely underestimates the requirement by missing out the testing needed to prepare for the second full reconstruction pass. However, accounting for this by adding an extra reconstruction pass adds an extra 5% total T1+T2 computing load.

Storage

- Applying the ALICE spreadsheet model for storage in 2009, there is no effect on disk storage requirements (the calculation assumes all derived data is generated in the same year as its parent real data).
- There is a small reduction in mass storage, as expected with no 2008 data to store. Our estimates for mass storage at (CERN, T1 ext, T2 ext) change from (7.7, 10.6, 0) PB to (7.1, 10.1, 0) PB.

Our simplified application of the ALICE computing model leads to the present assessment. While most of the numbers we found agree (within errors) with the request, the disk requirement* for T2s more than doubles the original request (whose

fulfilment already looked difficult). We recommend that the experiment reviews the computing model to deal with this fact.

2009 (revised)

Resource	CERN	T1 ext	T2 ext	Total	change	Sep 07 request	change
CPU/MSI2k	10	12	21	**43**	0	43.8	-2%
Disk/PB	2.5	9.9	9.6*	**22.1**	0	15.3	44%
MS/PB	7.0	10.1	0	**17.1**	-7%	19.7	-13%

Scrutiny of the ATLAS Experiment Request

Overview

Recent experience has given ATLAS a good understanding of the strong and weak points of their computing model. Starting with throughput and functional tests during CCRC08, and completing with several 'full dress rehearsals' (FDRs) in 2008, the model has been intensively tested. ATLAS is now able to assess its resource needs by considering practical experience and to analyse the impact of identified risks not only in the model itself but also in its implementation (both organisationally and in software).

The ATLAS computing model presented in the TDR was optimistic with respect to event sizes, event data formats, the distribution model and the required resource capacity. There were no uncertainties assigned to any of the input parameters. At this time some elements of the model should be reassessed. A proper risk analysis is the only reliable way of planning for future contingencies.

The 'FDRs' and CCRC08 have lead to a more realistic scenario for storage and ATLAS is actively addressing this even within their original resource envelope. To deal with the change in event sizes and with the storage needs of new formats that began to proliferate only after the TDR was completed ATLAS has until recently reassigned current pledged allocations to different functions but in doing so they risk sacrificing physics accuracy by reducing the fraction of simulated events relative to real events.

ATLAS has seen a proliferation of new event formats, and choices should be made as to which format is going to be primarily used for calibration, reconstruction and physics analysis. Storing the same data in several formats (and then in many copies of each) is wasteful given the constrained resources. A management-level decision to rely on either AOD or on the ensemble of physics DPDs is recommended to optimize resource usage.

Unless ATLAS makes a number of important and difficult choices (see recommendations below), the resource estimates by the CRSG for ATLAS in 2009 indicate that the computing needs for 2010 and beyond may be hard to materialize. An actual assessment of the ATLAS resource usage once data taking has started will give a more accurate indication of actual need in 2010 and beyond as it can take into account which risks, if any, actually materialised in real data taking operations.

To assess the ATLAS resource request for 2009, we obtained from ATLAS a simplified version of their model, which although not giving exactly the same values as the full model is accurate enough and does clarify better the relationships. The numbers presented below are obtained from this model.

Scenario of LHC operations assumed by this experiment

The last request by ATLAS, dated 14 August 2008, prior to the incident in sector 3-4, was submitted to us while the scrutiny process was already well under way. In this request, the parameters for 2008 assumed 3 x 10^6 live seconds, and in 2009 6 x 10^6 seconds, different from our standard assumptions common to all experiments.

While live seconds and luminosity assumed in the ATLAS resource request are different from the CRSG baseline assumption, the first-order dependency of the resource requirement on the live seconds is linear, whereas changes to the luminosity to first order do not affect resource requirements. Prima facie, it would appear that the ATLAS need for storage of 2009 LHC data (using the CRSG assumptions) may actually be larger than requested. However, the recent cancellation of 2008 running represents a decrease in ATLAS requirements which almost exactly compensates their use of a too small 2009 live time (according to our assumptions).

Since the Atlas scrutiny was the last to be consolidated, the effect of the new operating parameters (no live seconds in 2008 and 9x10^6 seconds pp in 2009) could be included in this analysis directly. In our analysis below we consider the ATLAS request in light of the live time values assumed by the CRSG (both pre and post-quench).

Storage requirements by site

The CRSG believes the ATLAS request is representative of their real need at the T0. The re-emphasis to the T0 is dictated by their computing model. The data buffer has about 60TB for 5 days of RAW input, and about 300TB for automated calibration sets in 2009. Throughput to disk is a limiting factor and ATLAS requests new additional disk servers. These servers and associated buffers are critical for maintaining data throughput to the T1s and without these transport of data to the T1s is not possible. These transport server buffers account for 45% of the requested disk capacity at the T0.

The CRSG 2009 estimates shown below purposely do not account for any possible cosmic ray data storage for 2009 while they do allow ATLAS to continue to store their earlier acquired data. ATLAS has indicated that, once data taking starts, cosmics will be gradually removed from disk to make place for real data.

2009

Tier 0 Disk (TB)	
ATLAS Request	650
CRSG Estimate (not affected by recent quench)	**650**

Tape requests for the T0 are cumulative. The CRSG estimate shown below attempts to account for the difference in 2009 live time guidance and the value assumed by ATLAS in their latest request. Unless otherwise stated, ATLAS request refers to the one submitted on 14 August 2008.

2009

Tier 0 Tape (TB)	
ATLAS request	8557
CRSG Estimate (pre-quench)	10395
CRSG Estimate (post-quench)	**8557**

The CRSG original (pre-quench) estimate assumed that all 2008 data will be stored (5282 TB) and a live time for LHC of 0.9 E07 sec, which is 1.5X the value assumed by ATLAS in their recent request. The post quench estimate assumes zero 2008 LHC events, and is estimated by scaling the ATLAS-provided 2009 LHC event estimates

back to the 2008 previously expected live time and then subtracting this 2008 event estimate from the 2008 component of the cumulative ATLAS 2009 request. It allows for an estimated 3645 TB of extant 2008 data to continue to be stored.

The CAF is dedicated to non-automated calibration, alignment and monitoring tasks. ATLAS has found the requirements to be greater than first planned. The buffer must hold 20% of a full ESD set (data will cycle through) and a full AOD set. There must also be space for the current version of 10% of the RAW. These reference sets at the CAF support high-priority analysis/algorithmic development within the detector and performance groups, such as analysis and verification immediately after run start in order to assess the current run parameters. This has required not only the provisioning of both of disk capacity for the DPD and AOD, and a disk-only service for intermediate or temporary results from calibration/alignment procedures, and for studies of detector performance based on datasets hosted on the CAF. The latter is sized at 1.5TB each for 300 active users in these dedicated groups. There is a corresponding growth in CPU requirement to allow these files to be produced and analysed. In particular, the CERN capacity is not intended for CERN-based people to do physics analysis, but for ATLAS physicists who may be in many sites to do time critical studies using a non-distributed system.

We feel that the original ATLAS model was skewed to the T1s, without sufficiently accounting for the need for near-real time calibration where the first pass analyses need to be at CERN.

2009

CAF Disk (TB)			
	ATLAS request	CRSG estimate (pre-quench)	**CRSG estimate (post-quench)**
Raw	274	274	**274**
ESD (inc. buffer)	886	1201	**886**
AOD+TAG (inc. buffer)	1058	1401	**1058**
Calib triggers	350	476	**350**
User/scratch	736	736	**736**
TOTALS	**3304**	**4087**	**3304**

In the previous table, the post quench estimate assumes zero 2008 LHC events, and is estimated by scaling the ATLAS-provided 2009 LHC event estimates back to the 2008 previously expected live time and then subtracting this 2008 event estimate from the 2008 component of the cumulative ATLAS 2009 request.

2009

CAF Tape (TB)			
	ATLAS request	CRSG estimate (pre-quench)	**CRSG estimate (post-quench)**
ESD	403	597	**597**
AOD+TAG	434	624	**624**
Calib	102	118	**102**
DPD	200	263	**200**
TOTALS	**1139**	**1602**	**1523**

The CRSG pre-quench estimate assumed that all 2008 will be stored (214 TB) and a live time for LHC of 0.9 E07 sec, which is 1.5X the value assumed by ATLAS in their recent request. The post quench estimate assumes zero 2008 LHC events, and is estimated by scaling the ATLAS-provided 2009 LHC event estimates back to the 2008 previously expected live time and then subtracting this 2008 event estimate

from the 2008 component of the cumulative ATLAS 2009 request. It allows for an estimated 135 TB of extant 2008 data to continue to be stored.

The resource requirements for both T1 and T2 have not changed substantially from earlier requests and, as mentioned earlier, reflect a commitment by ATLAS to live within these constraints outside CERN. However, as discussed above, ATLAS's assumption of 2/3X the expected LHC live time implies that data volumes from LHC data, exclusive of any cosmic ray data, will be 50% larger than they predict. The lack of 2008 data, however, cancels out this effect. This is noted in the tables below.

2009

Tier 1 Disk (TB)			
	ATLAS request	CRSG estimate (pre-quench)	**CRSG estimate (post-quench)**
RAW	1800	2138	**1800**
Proc (ESD, AOD,DPD)	14152	14152	**14152**
Buffer (new re-proc)	3076	3076	**3076**
User	2366	2366	**2366**
TOTALS	**21394**	**21732**	**21394**

The CRSG pre-quench estimate assumed that all 2008 data will be stored (10544 TB) and a live time for LHC of 0.9 E07 sec, which is 1.5X the value assumed by ATLAS in their recent request. The post quench estimate assumes zero 2008 LHC events, and is estimated by scaling the ATLAS-provided 2009 LHC event estimates back to the 2008 previously expected live time and then subtracting this 2008 event estimate from the 2008 component of the cumulative ATLAS 2009 request. It allows for an estimated 5119 TB of extant 2008 data to continue to be stored.

2009

Tier1 Tape (TB)			
	ATLAS request	CRSG estimate (pre-quench)	**CRSG estimate (post-quench)**
RAW	6200	7300	**6200**
Proc (ESD, AOD,DPD)	8850	8850	**8850**
TOTALS	**15050**	**16150**	**15050**

The post quench estimate assumes zero 2008 LHC events, and is estimated by scaling the ATLAS-provided 2009 LHC event estimates back to the 2008 previously expected live time and then subtracting this 2008 event estimate from the 2008 component of the cumulative ATLAS 2009 request. It allows for an estimated 5825 TB of extant 2008 data to continue to be stored on tape at the Tier1.

2009

Tier 2 Disk (TB)			
	ATLAS request	CRSG estimate (pre-quench)	**CRSG estimate (post-quench)**
RAW	579	869	**869**
ESD	1719	1719	**1719**
AOD+TAG	6274	6274	**6274**
User/DPD	5783	5783	**5783**
TOTALS	**14355**	**14645**	**14645**

As in previous tables, the post quench estimate assumes zero 2008 LHC events, and is estimated by scaling the ATLAS-provided 2009 LHC event estimates back to the 2008 previously expected live time and then subtracting this 2008 event estimate from the 2008 component of the cumulative ATLAS 2009 request.

Figures for the requested storage are consistent with the combination of event size, expected rate and number of copies. The ATLAS model incorporates efficiency factors expected for locally managed storage resources. However, the storage is managed in a distributed fashion using several highly complex pieces of storage and indexing and meta-data information that is independent from the physically stored objects. Each step introduced the possibility of inconsistencies, and inconsistency means making the data inaccessible. This results in inaccessible 'dark' data being resident in storage. The continued existence of software bugs should be acknowledged. It is realistic to assume that dark data will remain, at least at the ~ 10% level, which should be taken into account.

The most striking difference is the resource request for the T0 and the CAF in 2009 and the 'baseline' added to the 2008 tape usage at the T0. This reflects the new event sizes and the need to keep data available as the detector is being characterized. Current tape usage at the T0 is a fact that has to be acknowledged as a reality.

Monte Carlo requirements

The ATLAS computing model as presented in the TDR assumes a fully-simulated to real-data ratio of 20%, and this was entirely determined by CPU resource availability in the T2s. This is based on experience from the D0 experiment but is still significantly less than the simulated-to-real ratio used by comparable experiments.

The updated fraction of full MC generation of only 15% of the real-event fraction is sub-marginal, and may inhibit effective calibration in 2009. Ways to improve MC generation capacity (including more opportunistic use at the T3 level and below, or any other means) should be pursued.

Tier-0 and CAF CPU request and assessment

The CAF is used for initial calibration using actual events, and the output of the calibration at the CAF is used in the initial T0 processing when generating the ESD, AOD and other primary products before them being distributed to the T1s.

The CAF and T0 have to keep up with the incoming data rate from the detector at 200Hz during pp data taking. In AA running, events accumulate and the ensuing shut-down period is used to process the accumulated back log.

As pointed out above, the CRSG feels that until recently T0 and CAF resource requests remained consistently too low given the accrued experience within ATLAS.

The model expects a 2-day cycle for calibration and the storage and CPU requirements at the T0/CAF are derived from this. Based on the FDR experience, this turn-around time is too optimistic, and additional resources are needed either to speed up CAF processing, or to accommodate a longer pipeline. The increased CPU cost of the algorithms for calibration and first-pass ESD and AOD generation have a direct impact on the T0 and CAF CPU needs, required for near-real time validation and calibration tasks (within minutes of the start of each run). In addition, the CAF CPU has to accommodate these temporary alignment and calibration data and their generation and CPU here follows the growth in data volume at the CAF.

At the same time, disk resources located at CERN are increasingly needed for T0 use as experience has shown that, in order to attain the required throughput to the T1s, more data movement servers with local disk cache are needed to sustain the transfer rate (as much as 300 TByte extra is needed for this purpose).

The following tables summarize the ATLAS CPU request to CERN for the T0 and CAF and include the CRSG estimates based on the LHC 2009 live time guidance and assuming that no cosmic ray data are processed.

2009

Tier 0 CPU (kSI2K)			
	ATLAS request	CRSG estimate (pre-quench)	**CRSG estimate (post-quench)**
Processing	7058	9263	**7058**
Automated Calibration	529	794	**529**
Totals	**7587**	**10057**	**7587**

2009

CAF CPU (kSI2K)			
	ATLAS request	CRSG estimate (pre-quench)	**CRSG estimate (post-quench)**
Calibration	2169	2545	**2169**
Detector performance	3614	3614	**3614**
TOTALS	**5783**	**6159**	**5783**

The pre-quench processing requirements for 2009 assumed minimal cosmic ray analysis and assume the 2008 requirements are needed in 2009 to reprocess previous data. The CRSG estimate for 2009 assumes a live time for LHC of 0.9 E07 sec, which is 1.5 X the value assumed by ATLAS in their recent August 2008 request. The post quench estimate assumes zero 2008 LHC events, and is estimated by scaling the ATLAS-provided 2009 LHC event estimates and then subtracting these from the ATLAS 2008 request.

CPU usage at the T1s and T2s

ATLAS has realized that it will need to use Tier 1 capacity for simulation in the early years and the data management and work flows have been changed to accommodate this. This is necessitated by the increased time for full Geant4 simulation.

2009

Tier 1 CPU (kSI2K)			
	ATLAS request	CRSG estimate (pre-quench)	**CRSG estimate (post-quench)**
Reprocessing	12060	15325	**12060**
Simulation	5320	6690	**5320**
Group productions	11481	11481	**11481**
Calibration	530	795	**530**
TOTALS	**29391**	**34291**	**29391**

The capacity at the T2s is now insufficient to deal with the increased simulation times and off-loading this load to the T1s may impact the activities that should have been performed at the T1s for calibration and organised analysis, but by re-balancing the request the T2 load remains largely within the envelope.

2009

Tier 2 CPU (kSI2K)			
	ATLAS request	CRSG estimate (pre-quench)	**CRSG estimate (post-quench)**
Simulation	16575	19641	**16575**
Group/det	6780	6780	**6780**
Analysis	6966	6966	**6966**
TOTALS	**30321**	**33387**	**30321**

Observations and recommendations

ATLAS has submitted a request for a significant increment to their T0 and CAF resources while keeping their T1 and T2 resources requirements roughly constant since the TDR. Their use of the GRID model has evolved as they discovered the limitations to their computing model. The CRSG understands that the original ATLAS computing model was skewed more heavily to the T1s whereas near real time verification, within minutes from the start of a run, must be done at CERN.

In addressing their issues, ATLAS seems to blur the distinction between T0 and CAF resources, reallocating them to address shortcomings. A case in point is the late realization that throughput from the T0 to the rest of the grid hierarchy is limited by storage buffers and CPU capacity. This limitation appears to play a significant role in ATLAS's effort to place more emphasis on the use of CERN resources. The CAF is slated to perform near-real time priority physics validation and calibration tasks.

In summary, the key recommendations to be made are

- The bare minimum level of essential tasks that must be done at CERN should be determined and limitations enforced. The de-emphasis of the T1 and T2 roles relative to the CERN role is cause for concern for the reasons discussed above. However, the CRSG understands that the original ATLAS computing model was skewed to T1s whereas the bulk of the first pass analyses must be done at CERN. The reemphasis of CERN resources appears to be dictated by their experience. Nonetheless, ATLAS should consider how it apportions tasks and priorities among the grid hierarchy. It should consider further offloading CERN onto the T1s and T2s wherever possible, and strictly enforce the policy that the CERN capacity is not intended for CERN based people to do 'regular' analysis. Accommodation of extra, unexpected capacity should also be requested of the other grid elements.

- Event sizes have grown and event formats have proliferated, exacerbating the data volume challenges. ATLAS should take a hard look at possibly redundant utilization of different formats by different groups for essentially similar purposes. A management-level decision to rely on either AOD or on the ensemble of physics DPDs is recommended to optimize resource usage.

- The unexpected increase in simulation CPU costs needs to be addressed. The increased resources and diminished volume of Monte Carlo events which can be stored represent risks to ATLAS's ability to perform its key physics analyses efficiently within its resource limitations.

The effects of the September 19th events on the Atlas are fully consolidated in all estimates given above, indicated as the CRSG 'post-quench' or 'revised' estimate. Our recommendations are summarized in the following table along with the changes with respect to the 'historic' request (September 2007)

2009 (revised)

RESOURCE	T0	CAF	T1 ext	T2 ext	TOTAL	Sept 07 request	Change
CPU/kSI2K	7587	5783	29391	30321	**73082**	62020	+18%
Disk/TB	650	3304	21394	14645	**39993**	36300	+10%
MS/TB	8557	1523	15050	-	**25130**	22000	+14%

Scrutiny of the CMS Experiment Request

Overview

This section summarizes the outcome of two meetings and email discussions we had with the CMS computing management between May and August 2008. Subsequent to our discussions with CMS we produced an independent and significantly simplified spreadsheet that mirrors the CMS computing model. We find good agreement between this independent assessment and the CMS computing resource requests, with a couple of exceptions – noted below. We have further modified the inputs of our model to conform to the CRSG understanding of the LHC running scenarios in 2008 and 2009. This results in our revised estimates of the likely CMS computing needs. Again in all but a couple of cases these estimates are consistent (at the 10% level) with the CMS computing requests. As a warning, we indicate that except for the largest discrepancies, the differences may be due to the simplified model for the scrutiny.

CMS has made good progress in understanding the realities associated with their computing model. In particular their recent CSA08 computing exercise and the processing of first data from the CMS tests on the surface in the Fall of '07 and more recently from their zero-field run underground has confirmed that the computing resources and analysis model they foresee using when the first LHC data becomes available is viable. At the same time, nothing they have learned has led them to conclude that the resources anticipated in their computing model could/should be scaled back at this time. Thus we were convinced that they should stick with their 2009 computing resource estimates to be ready for the first full year of LHC running. The first encounter with proton-proton collision data will be crucial to refining estimates for 2010 and beyond. Given the resource acquisition cycle it seems prudent to scrutinise CMS's use of resources in the spring of 2009 to inform the Tier1 (and other) computing centre resource acquisition plans for calendar year 2010. A more detailed description of the analysis described below can be found in the full report on CMS to the CRSG.

CRSG scrutiny of resource requests

We have used the running time, event-sizes and data-formats in the CMS computing model to reproduce, at a simple level, the CMS computing requirements following the different assumptions about the luminosity profiles agreed by the CRSG.

Storage

The subdivision of the storage resources among CERN, Tier1s and Tier2s for 2008 and 2009 are given in the following table, for the default scenario of the LHC live time. A disk space efficiency factor of 70% has been taken into account. An overlap of 10% between the primary datasets has also been taken into account.

CERN Tape (Tbytes)	**2008**	**2009**
RAW, RECO, AOD *(incl calibrations)*	3870	8370
Total scrutiny	**3870**	**8370**
Total requested	*5300*	*9300*

Tier1 Tape (Tbytes)	**2008**	**2009**
RAW	1350	3300
Proc (RECO, AOD)	4770	13200
Total scrutiny	**6120**	**16500**
Total requested	*9800*	*15000*

T0 Disk (Tbytes)	2008	2009
Buffer for RAW data	285	285
Total scrutiny	**285**	**285**
Total requested	*400*	*200*

CAF Disk (Tbytes)	2008	2009
RAW	0	0
RECO	1285	2060
AOD+TAG (incl. buffer)	386	520
Calibration triggers		
User/scratch		
Total scrutiny	**1670**	**2580**
Total Requested	*1800*	*2300*

Tier1 Disk (Tbytes)	2008	2009
RAW	0	0
RECO	5650	5610
AOD	1450	1580
User	0	0
Total scrutiny	**7010**	**7200**
Total Requested	*7200*	*9700*

Tier2 Disk (Tbytes)	2008	2009
RAW	0	0
RECO	5660	0
AOD	1450	5300
User		
Total scrutiny	**7010**	**5300**
Total Requested	*5100*	*5700*

Our simplified analysis of the consequences of the inputs we have used to construct this model supports the CMS requests, despite the fact that we have not been able to fully assess the cumulative effects beyond 2009. Given the small size of the 2008 LHC dataset and the transient nature of the cosmic-ray/commissioning data that will dominate this period we believe our model is sufficient. The Tier0 and Tier1 resources we compute are within 10% of those requested by CMS. We believe the uncertainties in our model are at least this large. Our model (and probably CMS's) is incomplete for the Tier1 and Tier2 storage requirements and so it is not surprising that we find the largest discrepancies there.

CPU

We have used as input the CMS estimates of CPU/event and the dataset sizes to reproduce, at the simplest level, the CMS computing requirements under the CRSG assumed luminosity profiles to arrive at their CPU requirements at the various Tiers for 2008 and 2009. The Tier0 CPU requirements we compute are significantly higher than foreseen by CMS. The Tier2 CPU requirements we compute for 2009 is 40% lower than the CMS request.

Tier0 CPU (kSI2K)	2008	2009
Processing	11000	14700
Calibration	0	0
Scrutiny: Total	**11000**	**14700**
Total Requested	*5300*	*9800*

CAF CPU (kSI2K)	2008	2009
Calibration, etc.	2000	1700
Groups	0	0
Scrutiny: Total	**2000**	**1700**
Total Requested	*2100*	*3900*

From our analysis we conclude that CMS has under-estimated their Tier0 requirements for 2008. These appear to be driven by the reconstruction of raw data from the experiment, rather than calibration (and CERN-based analysis) activities on the CAF. By 2009 our understanding is that the discrepancy between request and requirements will all but disappear. Although we do not report on our findings for 2010 in this year's report to the C-RRB we note that the significant under-estimate in Tier0 computing seen here is exacerbated in subsequent years as the luminosity ramps up and the event reconstruction times continue.

The Tier1s are responsible for re-processings of the data as reconstruction code matures as well as the skimming of data-sets for user analysis. The following table summarises our understanding of CMS's Tier1 needs in 2008 and 2009.

Tier1 CPU (kSI2K)	**2008**	**2009**
Reprocessing	4240	7960
Simulation	3180	5970
Group productions	1670	2500
Calibration		
Scrutiny: Total	***9090***	**16430**
Total Requested	*9600*	*16300*

The CMS computing model foresees harnessing an amount of Tier1 computing power that is equal in size to their needs at CERN. Since almost all of the computing foreseen at the Tier1s is of the centrally organised variety it is, perhaps, not surprising that we were able to capture the CPU needs here in our model. At the Tier1s the CPU required is dominated by the centralised re-reconstruction of the CMS dataset once improved analysis code and calibrations become available. Once again, one worrying sign is that, beyond 2009, as the CMS event reconstruction is projected to grow to 125s per event (from 25s per event) as the LHC luminosity approaches its design, we foresee a significant shortfall of Tier1 CPU. This should be watched over the course of 2009 and every attempt to should be made to mitigate the inevitable growth that will occur as the CMS events become more complex.

In our attempt to capture the CMS computing model in our simplified spread-sheet we were successful at reproducing the centrally organised activities (1st pass processing and calibrations on the Tier0 and re-processing/skimming on the Tier1/Tier2). At the Tier2s the CPU requirements are driven by the need to generate centrally defined MC datasets. The CMS computing model still foresees three MC events will be produced for every four data events (down from 100%). These will be produced with the full GEANT model of CMS. This accounts for half of the Tier2 CPU required and our simple version of the CMS computing model captures this. Our scrutiny of the CMS scheduled skimming/analysis activities only amount to 15% of their requested Tier2 CPU. User driven analysis activities (ntuple production and thinning, final event selection and fits, estimate of systematic uncertainties) account for the remainder, and that the sum of scheduled and chaotic analysis activities is comparable to the resources required for MC production. While we can't currently substantiate the remaining Tier2 resource requirements for 2009, we do not think their requests are unreasonable. We accept that significant additional resources may be necessary and caution that it may be difficult for CMS to marshal all of their pledged Tier2 resources (at least initially in 2008 and 2009).

Tier2 CPU (kSI2K)	**2008**	**2009**
Simulation	5730	8590
Group/det (scheduled)	955	2390
Analysis (scheduled + chaotic)	4720	6290
Scrutiny: Total	**11405**	**17270**
Total Requested	*13400*	*28100*

Conclusions and Recommendations

The CMS computing model appears to be viable and has successfully weathered the various simulated test campaigns. Over the spring of 2008 the CMS computing group has simulated, reconstructed and distributed mock-data sets to all of their Tier1 and a

significant fraction of their Tier2 sites. At the same time they have absorbed, reconstructed and distributed data-taken at Point5 in their initial configuration at zero magnetic field – albeit without data from their tracker – the largest source or raw and re-constructed data once LHC collision data becomes available.

We see no reason to doubt that the CMS computing framework (model as hardware resources) will survive first contact with LHC collision data. Our scrutiny did not uncover un-warranted use of computing resources that would lead us to recommend a significant scaling-back of computing resources being pledged to CMS. Their anticipated use of a "local Tier1" at CERN (their CAF) for first pass calibrations and express-stream (re)-reconstruction appears well justified given the relatively limited experience available in HEP with the use of GRID-style computing for data. We cannot account for the relatively small request CMS makes for Tier0 CPU resources. This is one of the simplest calculations in our model. It does not depend on the live-time assumed (Tier0 reconstruction must keep up with data-taking in real-time). We suspect their model might not have been updated to account for the factor of 2 (1.3) increase in trigger rate expected for 2008 (2009), respectively, relative to the CMS computing TDR. We also note that our estimates of Tier2 disk needs are larger than the CMS requests. Once again we trace this to their stated goal of providing access to the RECO data, at the Tier2s, for the initial data-taking period. This is something we support but, in our model, it results in a larger call on Tier2 disk resources in the first year of data-taking. Our relative lack of experience with non-centrally managed computing on the GRID makes it difficult (essentially impossible, for us) to properly scrutinise the analysis tasks that will go on at the Tier2s. We have very little experience with distributed analysis, on the scale proposed here, in existing HEP experiments. As a result we believe our scrutiny numbers for Tier2 CPU are probably an underestimate, though we cannot quantify how much additional resources will be required at this time. This should be one of the primary questions to address in 2009 once the first real analyses have been performed. It would be appropriate to re-assess the resources required at that time. The table below summarizes the resources we recommend to be warranted to CMS for successful data taking and subsequent analysis in 2009.

2008 Summary

Resource	Tier0	CAF	Tier1	Tier2	Total
CPU (kSI2k)	11000	2000	9090	11405	**33495**
Disk (TB)	285	1670	7010	7010	**15975**
Tape (TB)	3870		6120	-	**9990**

2009 Summary

Resource	Tier0	CAF	Tier1	Tier2	Total
CPU (kSI2k)	14700	1700	16430	17270	**50100**
Disk (TB)	285	2580	7200	5300	**15365**
Tape (TB)	8370		16500	-	**24870**

Note added in the absence of the 2008 running

With the disappearance of the 2008 running we have attempted to revise our estimates of the resources required for 2009. We have looked at areas where the 2008 data was expected to persist into 2009 (and beyond) and revised our estimates accordingly. Although we have reduced the collider running time to 0 we still include one month (1.5×10^6s) of cosmic-ray data/calibrations processed through the full CMS chain. We re-examine the impact this, reduced, 2008 data has on the cumulative resources required by CMS for 2009. Beyond the propagation, through our scrutiny model, of the smaller data-sets there are two additional complications. Firstly, we assume that the event sizes at startup in spring 2009 will be the same as the ones foreseen for autumn 2008 in the original plan. Although we believe that a sizeable reduction in data sizes can be

accomplished even with cosmic data in the absence of real beam data, we nevertheless allow for some contingency. Secondly, our model foresaw/allowed the distribution of one copy of the full 2008 RECO dataset to the Tier2's, to provide the collaboration and working groups with better access to the first data allowing them to converge more quickly on their final reconstruction algorithms and AOD data format. We still believe this will be a necessary step and thus include the provision to distribution 1/3 of the 2009 data, in RECO format, to the Tier2s. The impact of this change is reflected in the Tier2 disk and CPU estimates summarised below.

We note that the CERN/Tier0 CPU resources are un-changed since they must provide real-time throughput as the data is taken and the live time assumptions for 2009 are un-changed. Since tapes are no longer needed to archive the 2008 data, the corresponding resources at CERN and Tier1 are reduced by about 10%. The disk requirements are practically unchanged, and are actually increased at CERN. This is due to the event sizes in 2009, which stay at the same level as in 2008. The Tier1s see some CPU reduction since they are no longer required to serve or reprocess (much) of the 2008 data. The Tier2 resources see correspondingly smaller reductions. Roughly half of their mission is to provide MC. We did not change our MC assumptions because of the absence of 2008 data.

2009 (revised)

Resource	Location	Original	Revised	Change	Sep 07 request	Change
CPU (MSI2k)	Tier0/CAF	16.4	**16.4**	-	58.1	-5%
	Tier1	16.4	**12.8**	-22%		
	Tier2	17.2	**15.5**	-9%		
Disk (PB)	Tier0/CAF	2.9	**3.4**	18%	17.9	-9%
	Tier1	7.4	**7.5**	1%		
	Tier2	5.3	**5.3**	-		
Tape (PB)	Tier0/CAF	8.4	**7.5**	-10%	24.3	-6%
	Tier1	16.5	**15.2**	-8%		

Scrutiny of the LHCb Experiment Requests

Overview

The LHCb experiment will collect data at a total rate of 2 kHz with data flows in four overarching categories:

- Exclusive b (specific B decay modes) at 200 Hz,
- Inclusive b (a B secondary vertex found) at 900 Hz,
- J/Psi channel (inclusive prompt J/Psi) at 600 Hz,
- D* channel at 300 Hz.

The first two data flows correspond to the signal and the main source of background, respectively. The last two as well as of interest for the physics itself are also used for calibration of the LHCb sub-detectors (muon detector and RICH detectors).

The trigger system is organized in two levels, the Level 0 (L0) and the High Level Trigger (HLT). The L0 trigger is used to reduce the bunch crossing rate of 40 MHz to 1.1 MHz, selecting events according to the highest p_T particles in the final state (muons, hadrons, gamma or electrons/positrons). L0 is implemented by means of custom electronics and can be configured to change the sharing of the bandwidth amongst the four data streams listed above. The HLT runs at 1.1 MHz as a software

trigger implemented on an online compute farm that runs the respective trigger selection algorithms. The HLT compute farm architecture is modular and scalable such that CPU power may be added in case of need. The HLT can be tuned to set the output rate as well as the sharing of the bandwidth amongst different data streams.

Offline reconstruction is performed in two steps. In a first step, RAW data are reconstructed into the intermediate rDST format. In a second step, rDST are selected and only a fraction of the data samples is converted into the final AOD (DST) format. rDST data are used to select relevant decay modes during the so-called selection or stripping phase, thus significantly reducing the final number of events for physics analysis by almost a factor of 10.

Reconstruction of RAW data into the rDST format, the following stripping step and the final reconstruction of the selected samples into the AOD format are performed in quasi real-time conditions at CERN and the 6 LHCb Tier-1 centres. A second pass reconstruction is planned during two months of accelerator shutdown making extensive use of the LHCb online farm. The Monte Carlo data are produced in the Tier-2 centres. The analysis is performed at CERN and the six Tier-1 centres, for which a full replication of data is foreseen (7 copies), whereas the full set of MC data reside at CERN and another 3 copies are distributed among the six Tier-1s (4 copies in total).

LHCb Requests

The total requirements of LHCb for disk, tape and CPU are presented in the following tables. Upon request of the CRSG they are based on a data taking time of 0.3 x 10^7s in 2008, 0.9 x 10^7s in 2009, and 10^7s in 2010 and beyond. These summary tables include the efficiency factors for CPU, disk and tape as agreed within the WLCG collaboration.

The following table set provides a breakdown of the total requirements in 2008 and 2009 for the different physics categories.

Total CPU power at all sites, including the LHCb online farm (MSi2k*year)	**2008**	**2009**
Stripping	0,4	1,4
Reconstruction	1,1	3,2
Monte Carlo	4,6	11,4
Analysis	0,5	1,8
Total	**6,6**	**17,8**

Total Disk at all sites (TB)	**2008**	**2009**
RAW	60	170
rDST	30	100
MC-AOD	370	940
AOD	490	1590
TAG	100	300
Analysis	110	400
Total	**1160**	**3500**

Total Tape at CERN and Tier-1s (TB)	2008	2009
RAW+rDST	660	2640
AOD + MC-AOD	440	1780
TAG	50	205
Total	**1150**	**4625**

The following tables provide a breakdown of the total requirements in 2008 and 2009 per site.

CPU (MSi2k*year)	2008	2009
Online farm	0,36	0,90
CERN T0 + T1	0,30	1,00
Tier1s	1,35	4,53
Tier2s	4,55	11,38
Total	**6,56**	**17,81**

Disk (TB)	2008	2009
Online farm	--	--
CERN T0 + T1	294	912
Tier1s	859	2566
Tier2s	9	23
Total	**1162**	**3501**

Tape (TB)	2008	2009
Online farm	--	--
CERN T0 + T1	489	1970
Tier1s	661	2656
Tier2s	--	--
Total	**1150**	**4626**

CRSG commentary and recommendations

The CRSG has analysed the computing model implemented by the LHCb experiment and concluded that the model is viable and solid. The successful recent tests have demonstrated the stability of the implemented solutions and suggest that the model will enable a successful data taking at the LHC start-up and beyond.

However, the CRSG recommends the careful reconsideration of the strategy for a full replication of the DST data at all Tier-1 centres and to review this model assumption in the present real conditions expected for 2009, such that a reduction of the replication factor could be envisaged in the future if allowed by the analysis flow and by the system stability.

LHCb plans to utilize the online farm for reprocessing within two months after accelerator shutdown. This requires fast access to large amounts of data from the online farm at the pit to the CERN Tier-0 storage located on the campus. According to LHCb the respective infrastructure has not yet been fully set up nor proven to provide the required bandwidth. The estimated compute power of the online farm for this reprocessing activity by the end of 2009 is quite substantial: 5,5 MSi2k or the equivalent of half a Tier-1 that serves for all four experiments. This compute power must be provided elsewhere if the above setup is not ready in time. Since testing this setup is hardly possible during accelerator runtime after spring 2009 the CRSG recommends that LHCb severely keeps track of the project plan for the implementation and testing of this local reprocessing setup but also starts to elaborate a fall back

solution in parallel. LHCb otherwise risks to not having enough CPU resources for the second pass reconstruction at the end of 2009.

In conclusion, the CRSG supports the requests of the LHCb experiment and recommends to the C-RRB the funding of these resources for a successful running in 2009 and beyond as requested. To summarize:

2009

Estimation of **total** resource requirements for LHCb (beam time 0.3×10^7 s in 2008 and 0.9×10^7 s in 2009)	
CPU (MSi2k)	**16,9***
Disk (TB)	**3501**
Tape (TB)	**4626**
(*) Note that the online farm which is only used within 2 months has been removed in this summary table in order to convert the CPU requirement into the usual MoU units of the installed capacity given in MSi2k.	

Note added in response to the absence of 2008 running

After the helium leak into sector 3-4 of the LHC tunnel and respective information that a restart of the accelerator complex has to be shifted to early spring 2009, the reviewers were requested to estimate the consequences for zero beam time (instead of the anticipated 3×10^6 seconds) in 2008 on the compute resources in 2009. The following table summarizes the results of a simulation in terms of total required CPU, disk and tape capacity respectively.

2009 (revised)

Estimation of **total** resource requirements for LHCb (zero beam time in 2008 and 0.9×10^7 s in 2009)		Change	Sep 07 request	Change
CPU (MSi2k)	**16.4***	-3%	17.4*	-6%
Disk (TB)	**3238**	-8%	3773	-14%
Tape (TB)	**3516**	-24%	5340	-34%
(*) Note that the online farm which is only used within 2 months has been removed in this summary table in order to convert the CPU requirement into the usual MoU units of the installed capacity given in MSi2k.				

A comparison with the summary tables for the original model above shows that

- The total CPU requirement in 2009 would reduce by about 0,5 MSi2k or 3% of the original request, 130 kSi2k at CERN and 65 kSi2k at each of the 6 Tier1s. The percentage impact is small because the CPU requirement is dominated by the Monte Carlo simulation as well as reconstruction and analysis of data from the current running year, i.e. 2009. The CPU power of a current dual CPU/quad core machine is of the order of 15 kSi2k, which allows for a rough estimate in terms of real hardware.

- The total disk requirement in 2009 would reduce by about 263 TB or 8% of the original request, 95 TB at CERN and 28 TB per Tier1. Tier-2 disk storage is not affected since Tier-2 centres are foreseen to produce only Monte Carlo data. The reduction in disk requirement comes from the fact that LHCb plans to keep on disk AOD, TAG and Analysis data of the previous year (i.e. 2008).

- The total tape requirement in 2009 would reduce by about 1110 TB or 24% of the original request, 470 TB at CERN and 107 TB per Tier-1. The percentage impact on tape is much larger than on disk because of missing RAW and rDST data from 2008 that would have been stored on tape under normal planned beam conditions.

It should be noted that the above estimates were done under the assumption of zero data from the detector and hence no analysis activities at all. This is certainly not realistic. Even without a beam LHCb has already done and will continue to do analysis of cosmic ray events etc. to study the detector in more detail, and certainly wishes as well to archive these scientific data to tape. However, respective data rates and data volumes are not contained in the LHCb model spread sheets and can hardly be estimated. We believe that they can be handled with the resources installed in 2008.

Still, the total reduction by about 1,1 PB in tape space for 2009 is alarming at a first glance. However, with the current tape technology this translates to roughly 590 tape cartridges at CERN and 135 cartridges per Tier1 (e.g. LTO-4, 800 GB/cartridge). The total cost for tape storage is roughly composed of 65% for the tape storage infrastructure (robotics, drives, switches) and 35% for tape media, and the Tier0 and Tier1 sites have already installed (or are currently about to procure/install further tape robots) with 6000 slots or more for April 2009 (communication within the WLCG Management Board). Furthermore, expansions of already existing tape robots are difficult without longer maintenance downtimes and are thus difficult to carry out during data taking. We thus recommend the funding agencies and sites to stick to the original plans concerning installation of the tape infrastructure. In view of the recent LHC incident with successive large reductions in tape requirements, buying tape media on demand, e.g. per quarter of a year, is certainly justified.