CERN-RRB-2010-052

Computing Resources Scrutiny Group

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

INTRODUCTION

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 2011, and the usage made of these resources in 2009. We also provide a first tentative estimate of the resources for 2012 which should be helpful for planning purposes. The latter will be revised in the April 2011 C-RRB.

The CRSG is an independent committee whose members are selected by the Funding Agencies represented in the C-RRB. The purpose of the CRSG is to inform the decisions of the Computing Resources Review Board (C-RRB) for the LHC experiments.

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-funding.

The starting point of the scrutiny is the resource request presented by the four experiments. 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 in the past of these resources and the accounting figures regarding usage and availability of the pledged resources

This year this report has more comprehensive contents than in previous years. For the first time there was real data to be taken and analyzed. Even though the total amount was lower than originally hoped for 2009, this has triggered a qualitative change in the scrutiny as the computing models are finally being put to a real test. The new scenario requires us to provide to the C-RRB, according to the CRSG mandate, a detailed account of the usage that the experiments have made of the computing resources and of the availability of the pledged resources.

The LHC schedule

The expected schedule of the LHC is an essential ingredient of the scrutiny even though a fraction of the computing needs do not scale with running time.

The expected beam time in 2010, 2011 and 2012 was subject to several revisions during the last months of 2009 and the early months of 2010. The final schedule was decided after the Chamonix meeting, in February 2010. It was agreed that, barring unforeseen circumstances,

the LHC should run for the best part of 2010 and 2011 (the estimate is 8 months per year), with only a relatively short break at the end of 2010 and beginning of 2011. It is expected to accumulate an integrated luminosity of 1 inverse fb. The energy has been decided to be 7 TeV (3.5 GeV + 3.5 GeV) for the whole period. About 10% of the time is expected to be dedicated to heavy ion (HI) physics.

After this long run a long shutdown would follow to enable the machine to reach the design energy in 2013 and the following years.

For the scrutiny the relevant quantity is the total number of seconds when the beam is declared to be stable and good for physics. This was the object of some amount of discussion as naturally there are many unknowns in a new accelerator, particularly of the complexity of the LHC. For this issue we thank the guidance and collaboration provided by Roger Bailey and, particularly, by the Director of Accelerators, Steve Myers. The chair of the CRSG was invited to attend a Machine Committee meeting prior to the LHC re-start in February where some of these issues were discussed.

As a consequence of these discussions the following estimation was agreed and shared by the WLCG Management Board:

Live time: 30 days/month = **720** hours

Folding in efficiencies 720 x 0.7 x 0.4 = 201.6 effective hours/month = **725760** sec/month

This leads to the following table that has been used by the four experiments when establishing their computing needs.

RRB year	RRB year start	RRB year end	Months (max) data taking	Total live time (in Ms)	pp	HI
2009		**May '10**	**3**	**2.2**	**2.2**	**0**
2010	**June '10**	**March '11**	**8**	**5.8**	**5.1**	**0.7**
2011	**April '11**	**March '12**	**8**	**5.8**	**5.1**	**0.7**
2012	**April '12**	**March '13**	**0**	**0**	**0**	**0**
TOTAL			**19**	**13.8**	**12.4**	**1.4**

Interactions with the experiments

After the Chamonix meeting in early 2010, there was a first round of interaction with the experiments in the course of a joint LHCC-CRSG meeting on February 16th. The four experiments presented a summary of the utilization they made of the resources during the year 2009. They also presented preliminary requests for 2011 and a tentative prospective for 2012 assuming that the presently approved schedule is adhered to without further changes.

The experiments were reminded, following the recommendations of last year's report to the C-RRB of the need to submit a detailed account of the usage of the resources by March 1st., which they did with reasonable diligence.

They were also asked, again following the indications of last year's report, to submit their request for evaluation by the March 1st deadline. Unfortunately, we received some of the requests as late as March 17th, with some complementary documents arriving even later. The CRSG recognizes the difficulties encountered by the experiments in planning their computing resources, but the short time available to this committee then makes our task much more difficult and is the ultimate justification as to why this report is submitted to the C-RRB with such short notice, for which we apologize.

The experiments requests contained this time a limited number of changes with respect to previous iterations of the scrutiny procedure. This is an indication of the maturity of the computing models. Nevertheless some changes have been proposed by the experiments. Usually they have to do with data samples being larger than expected (but some decrease has also been noted in some cases as well as reductions in the processing and reprocessing times). Generally speaking we endorse the changes due to larger sizes and/or processing times as they are unavoidable in the starting period but we are of the opinion that there is room for optimization as more experience is gained. In any case, we ask the experiments to reach a compromise in the requested resources and to try to establish a reduction profile.

In other cases the revisions originate from changes in the data placement policy. These changes may have a substantial impact on the amount of disk requested and may represent changes that, if agreed, could make some of the computing models less sustainable. The CRSG believes that the existing computing models have largely proven their validity and sees no compelling reason for radical changes at this point without hard experimental evidence. We have therefore examined all requests along this direction more critically.

Following the reception of the requests, referees were assigned to the different experiments and a number of conversations and meetings took place. The experiments were provided ahead of the meetings with a set of questions in order to expedite the discussion.

As agreed with the ATLAS and CMS management in 2009 the scrutiny procedure for these two experiments is done by a common team of referees, using analogous techniques and methods. This procedure would ensure that a coherent set of principles is applied.

Generally speaking the interactions with the experiments are quite fluid and we thank the respective management for their openness and collaboration. However the CRSG is nevertheless concerned about the short time it was allowed to develop its scrutiny.

The CRSG recognizes the special characteristics of this combined run, extending over two years, and the uncertainties that remain in the current computing models and the need for redundancy/contingency. The situation for 2010 should indeed be regarded as somewhat exceptional, and to a large extent driven by the transient character of the start-up months, but we understand that commissioning activities will be largely over by 2011 and have scrutinized the resource requests in view of this.

The CRSG commits itself to undertake a revision of the present scrutiny ahead of the October 2010 C-RRB in the light of the usage of the resources up to that moment..

Interactions with the LHCC

In carrying out the scrutiny of the experiments requests the scope of this group is largely limited to the implementation of the respective computing models which are periodically reviewed by the LHCC.

The 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. Obviously there is a gray area where the respective competences of the LHCC and the CRSG overlap and it is not so clear what would represent a change of the computing model or what would just be a natural adaptation of it to the changing circumstances.

Over the last months a very fluid dialogue has been established between the CRSG and the LHCC including now joint meetings to review the experiments implementation of the models, computing usage, and the computing needs.

It should however be borne in mind that the nature of the scrutiny carried out by the CRSG is different to the one implemented by the LHCC reviews. For the scrutiny group, acting on behalf of the Funding Agencies, the budgetary cost of the resources made available to the experiments and their efficient use are relevant issues.

During 2010 no issues appeared for which we thought it was necessary to refer to the LHCC. However we would like the LHCC to keep monitoring the following points of concern:

- The event size has a very direct impact on the computing requirements. Experiments should make an effort to reduce the raw event size (and the size of all subsequent derived formats) and event processing time by establishing a reduction profile without unduly jeopardizing the physics. The experiments should also be prepared to make some compromises by revising their data distribution policy, reducing the number of copies stored in Tier 1 or Tier 2, for instance.
- The potential proliferation of different data formats serving the same purposes should be watched closely. Strong overlap between different primary data sets may also be a matter of concern.
- Care should be taken that the worldwide LCG resources are used as much as possible as there may be a tendency by collaborations to place heavier demands on CERN resources or suggest that a larger than originally planned part of their analysis should be done at CERN.
- Both CMS and ATLAS plan to reprocess at sustained rates around 400Hz – 500Hz in coming years. In view of the duty cycle of the LHC this may not be strictly necessary. Note that this requirement may have a large impact on cost.
- The CRSG encourages a close collaboration of the different Tier centers with the experiments to achieve an efficient and cost-effective access to data. Intelligent storage management requires close information exchange and coordination between the computing fabric and the experiment management tools. Only then concepts like pre-staging of data allow for a cost effective use of disk and tape while ensuring high CPU efficiencies. This type of sophisticated storage management will be especially needed in the area of end user analysis activities. Generally speaking, the CRSG encourages pre-staging from tape for centrally organized activities as a cost-effective measure.
- Different collaborations still estimate the impact of pile-up due to the foreseen running conditions differently and uniformity in this respect should be sought.

On the scrutiny process

The CRSG is now satisfied with the quality and quantity of the information provided by the experimental collaborations. The experiments requests are well documented.

The CRSG was provided on this occasion with simplified spreadsheets by the experiments that made the work of the referees easier. We thank the experimental collaborations for making this additional information available to us. These simplified spreadsheets agreed consistently with the ones made of our own, constructed from the information contained in the computing models and the written requests submitted to us. We note that some experiments request

resources that are systematically higher than the ones resulting from their simplified models or from the CRSG understanding of the respective computing models. The deviations are not always negligible., but unfortunately we have never been able to scrutinize in detail the origin of these differences as specific information is not provided. The recommendations provided here (sufficient to the best of our knowledge) are strictly based on the information made available to us.

For future reviews we insist that the following good practices be maintained:

- Requirements and respective models should be frozen during the review
- All changes to the models compared to the previous review should be well documented
- All documents should be provided sufficiently early to allow time for the review, a deadline for the revised requirements should be agreed upon well ahead of the final report deadline. For the upcoming October C-RRB meeting this deadline is **September 1st 2010**.

It is clear that the current implementation of the WLCG MoU should be agile enough to respond to the changing environment we face during the early days of the LHC. In the past a slow reaction has had a non-negligible cost for the funding agencies. The CRSG is committed to make recommendations to the C-RRB with the aim of optimizing the resources already invested and adjusting the future ones to the required computing needs.

Regarding the 2011 request, to the best of our knowledge the recommendations contained in this document should suffice to provide enough resources until 1st April 2012. However, the CRSG plans to closely follow during the coming months the implementation of the computing models and is prepared to submit a revised estimate during the October 2010 C-RRB meeting should it become necessary. This recommendation may go in the direction of recommending an increase in the resources for the second half or the RRB year 2011, or to recommend a shift of the 2012 pledged resources as it has been done in the past if it becomes apparent that there is an over-commitment of resources (we consider this last possibility unlikely under normal circumstances).

The experiments produced some very tentative estimates for 2012 (i.e. for the RRB year starting on April 1st 2012). These have not been fully scrutinized yet, but we should be able to make a more definite recommendation in the October 2010 C-RRB when experience gained once real data-taking becomes more routine should considerably reduce the remaining uncertainties, allowing better estimates beyond 2011.

General comments

It seems prudent to scrutinise closely the experiments' use of resources after the first months of data taking in 2010. The CRSG commits itself to provide such a report in the shortest delay which is feasible and in any case before the October 2010 meeting. In order for this to happen, the CRSG will require timely resource usage reports from all the experiments no later than **1st September 2010.**

To conclude this introduction we list some recommendations to the experiments that partly overlap with the points brought to the attention of the LHCC above

- The experiments should incorporate the running conditions into their models in a uniform way. At present there is still a discrepancy in the way that ATLAS and CMS treat the consequences of pile-up due to the foreseen long bunch crossings.
- 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. Reprocessing and simulation times also have a major impact and experiments should intensify their work in the direction of reducing these.

- We recommend the experiments make maximal use of the distributed resources in the GRID, thereby minimizing 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.
- As far as data distribution is concerned, we notice different strategies which try to optimize the total CPU power required to analyse data, ranging from the maximization of replicas among the computing centres to the usage of streamlined primary datasets. Each approach has its own advantage and disadvantages. We recommend that the experiments use the upcoming data taking period to determine which strategy optimizes physics output while keeping resource requirements at a reasonable level in the understanding that some of the tenets of the original computing models may become unsustainable.

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. There is no contingency for late delivery or failure to meet the pledges included in our estimates or for less than 100% availability of these resources.

Any shortage of CERN resources implies potential disruption of data taking; therefore, we advocate for a full support of CERN resources. Resources at Tier1s and Tier2s are crucial for physics output; any shortage of non-CERN (non-custodial) resources does not impact data samples, but slows down the physics productivity of the experiments.

On the CRSG membership

Two members of the CRSG left since the last C-RRB meeting. We have welcomed two new members: Martin Gasthuber from DESY who provisionally replaces Holger Martin and Bernd Panzer-Steidel from CERN who provisionally replaces Jürgen Knobloch, also from CERN. The appointment of these individuals has been made by the respective funding agencies but it is contingent on the approval of the C-RRB. Following the April 2010 C-RRB three new members should join the CRSG in order to comply with the mandate of the WLCG MoU.

Jonathan Flynn from the University Southampton has a rather serious accident and has been unable to actively participate in the present scrutiny. The members of the CRSG wish him a quick and complete recovery.

PART A

Scrutiny of the WLCG resources utilization in 2009

The re-commissioning of the LHC in 2009, following the incident on September 2008, was very successful but the total amount of LHC collision data delivered was small, only a percentage of the total physics data anticipated. In addition to the small amount of data that was considered good for physics the experiments accumulated a much larger amount of cosmics data, used for purposes of calibration and alignment of the detector systems, but the interest on this type of data sharply decreased as real data became available.

Not surprisingly, the experiments have not used all resources made available to them, but the usage of the WLCG resources has been both extensive and intensive. The tools and middleware appear now to be sufficiently mature to warrant a fruitful exploitation of the grid. This readiness combined with the enormous interest and enthusiasm in the first real LHC events, even if most of the data was really useful for commissioning only, triggered a noticeable increase in the use of the WLCG resources. This is particularly remarkable in the case of some of the collaborations that have shown a spectacular increase in the usage of the WLCG resources.

Sources

The information provided here comes from the WLCG accounting report for 2009

http://lcg.web.cern.ch/LCG/accounting/Tier1/2009/december-09/Master_accounting_summaries_December2009.pdf ,

the EGEE accounting portal at CESGA

http://www3.egee.cesga.es/gridsite/accounting/CESGA/tier1_view.html ,

and the reports that the experiments have provided to the CRSG with rather complete usage reports that are appended to this document. We thank the experimental collaborations for their cooperation.

The report refers, unless otherwise stated, to the calendar year 2009, from January to December. The CRSG believes that it is a good practice to be continued that the experiments report in **March 1st** about the usage made during the previous calendar year.

Overall usage

The following table describes the degree of usage of the different resources made available to the experimental collaborations by the WLCG

Resource	Site(s)	Used/Available
CPU	CERN	26%
	T1	69%
	T2	(see text below)
Disk	CERN	71%
	T1	81%
	T2	N/A
Tape	CERN	41%
	T1	64%

The figures for CPU usage correspond to a time average over the year 2009, obtained from averaging the monthly figures; those referring to disk or tape reflect the amount of the usage relative to the installed capacity at the end of the accounting period (December 2009).

While the installed CPU capacity at the Tier 2s is not centrally accounted yet, the CPU usage at the Tier 2s has actually amply surpassed the pledge in 2009 The CPU delivered by the Tier 2s to the experimental collaborations in 2009 was in average 168% of the pledged value. There are two reasons for this. One is due to the initial assumption that the efficiency associated to the Tier 2s is assumed to be 60% (as the Tier 2 centres should mainly run analysis programs) whereas it is actually much higher (see below). Another reason is revealed after an examination of the time profile of the accounting: resources installed were largely surpassing the pledge in the months from May to September because the deadline for 2009 was moved to autumn following the delay in LHC commissioning. The Tier 2s being more disconnected from the decision centres had no time to react to this postponement. This clearly reflects a malfunction of the decision and communication procedures.

The time profile of the installed disk capacity at the Tier 2s is not centrally reported and it is not available. The experimental collaborations have tools to known how much disk is available at a given time in a Tier 2, but no statistics are kept. Generally speaking the accounting tools for the Tier 2s are still not satisfactory and clearly more work on this is recommended.

The current percentages represent a noticeable improvement in an efficient use of the resources with respect to previous reports.

.

Usage by the individual experimental collaborations

Below we provide some tables, ordered alphabetically by experimental collaboration, that give a rather accurate picture of the experiments usage of the resources. Every table is followed by some comments. The data has been provided by the collaborations themselves and cross-checked whenever possible.

Note that the final request in 2009 was actually lower than the resources finally available, as explained later.

Resource	Site(s)	Used/Available
CPU	CERN	16% (*)
	T1	29% (*)
	T2	53% (**)
Disk	CERN	19%
	T1	29%
	T2	16% (*)
Tape	CERN	17% (*)
	T1	10% (**)

(*) We quote the used / pledged usage, as ALICE has not supplied percentage of usage to us but only absolute figures. Please see below to see the respective percentages of fulfilment of the pledges (which is 80-120%, depending on the resource)

(**) CPU and disk for Tier 2s is not completely accounted yet, the degree of fulfilment of the pledges in this case is less reliable than for CERN or Tier 1s.

Comments on the ALICE usage report

ALICE recorded a total 1M pp collision events with a total raw data size of 371 GB. Cosmics recorded in this period amount to a total raw data size of 85 TB, while calibration data amounts to 36 TB. Only raw collision data have been replicated 2 times (all together 3 copies including CERN's) in Tier 1's mass storage.

All raw cosmic and collisions data have been reconstructed once at the Tier 0 and several times (6 times) at the Tier 1s with continuously improved calibration and alignment parameters, many more times than contemplated in the computing model, in this initial period. The size of the ESD is 0.290 GB/event, a factor 10 larger than the value used in the computing model. As of today, the reconstruction process exceeds the 2 GB memory limit by about 50%, preventing ALICE to run jobs in several sites which enforce the 2 GB limit

The ESDs have been replicated 3 times and distributed in the Tier1s and Tier 2s disk.

The ALICE computing model assumes 2 copies of the ESD data at CERN and the Tier 1s plus a fraction on Tier 2 storage, but because of the small amount of data more copies were produced (an adaptation for the first LHC running period).

Analysis has been run indifferently on Tier 1s and Tier 2s where ESDs are stored. During data taking, a fraction of the events (about 10%) are reconstructed on the CAF.

Generally speaking ALICE appears to have made a use below average of the WLCG resources.

Resource	Site(s)	Used/Available
CPU	CERN	30%
	T1	97%
	T2	154% (*)
Disk	CERN	96%
	T1	76%
	T2	38% (**)
Tape	CERN	58%
	T1	39%

The figures for CPU usage correspond to a time average over the year 2009, those referring to disk or tape reflect the amount of the resource installed at the end of the accounting period (December 2009).

(*) Used/pledged. Available CPU at Tier 2s is not properly accounted.

(**) Snapshot on March 1st 2010. As explained elsewhere disk accounting at Tier 2s is not yet implemented.

Comments on the ATLAS usage report

ATLAS collected in 2009 170 TB of RAW data plus 850 TB of cosmic data. The number of collision candidates was close to 1M, more than half of them with the stable beam flag.

ATLAS used 2.2M HS06-days at CERN resulting from only having a short collision data-taking period at the end of the year. During LHC data-taking they were able to saturate the Tier-0 CPUs.

In several Tier 1 sites ATLAS has not used a significant fraction of the resources available to them. Notably CERN'sT0 and CAF, but in several other Tier1 sites the usage has fallen below the 90% line: US-BNL, UK-RAL, NL-NIKHEF/SARA and FR-CCIN2P3. According to ATLAS the brokering code (the code that distributes tasks to the sites where they can be executed) is still not perfect; it did happen from time to time that one or the other Tier-1 centre were entrusted with fewer jobs than they could have run. When this happened, its effects were more evident at larger centres than at smaller ones. Other sites have more than compensated this however as the total usage by ATLAS from the resources is close to 100%. ATLAS has made a good opportunistic use of idle CPU power in sites allowing it.

ATLAS distributed data in 2009 with more copies (termed 'secondary') than described in the computing model in order to ease access for detector analysis. These secondary datasets will be deleted when space is needed.

The resources made available to ATLAS were actually larger than they requested in September 2009 due to the shift in the commissioning of the LHC.

CMS

Resource	Site(s)	Used/Available
CPU	CERN	36%
	T1	43%
	T2	59%
Disk	CERN	60% (*)
	T1	82% (**)
	T2	40% (***)
Tape	CERN	47%
	T1	87%

The figures for CPU usage correspond to a time average over the year 2009, those referring to disk or tape reflect the amount of the resource installed at the end of the accounting period (December 2009).

(*) Average usage over the year: 72%

(**) Average usage over the year: 78%

(***) As indicated elsewhere accounting for Tier 2 disk is not available.

Comments on the CMS report

According to their report, during 2009 CMS collected over 0.5M events declared good for physics, although the total numbed of events recorded was about 10 times larger.

The resources made available to CMS were actually larger than they requested. For a comparison the usage/requested fractions were: T2 disk: 96%; CERN tape: 64% and T1 tape 120%. This reveals an over commitment of computing resources, tape notably.

As it will be shown below, CMS did 35% of their CERN + Tier 1 computing at CERN. The CMS request for processing in 2009 was 44.7 kHS06 for CERN and 46 for the Tier-1s. This means that nearly half of the computing to be done at CERN + Tier 1 was planned to be at CERN only. Their computing model is relatively centralized at CERN. It allows, for instance qualified individual users to submit jobs to the CAF with a limit of 100 running jobs per individual. In spite of this, CMS reports a relative usage of 13% of the CAF resources indicating the desire of the collaboration to commission the distributed infrastructure for analysis. Note that in the CMS TDR the label CAF includes the equivalent of a sizeable Tier 2 data center at CERN.

CMS reports a temporary modification of their file distribution policy: In the CMS computing model only 10% of the simulated events are expected to be staged on disk, but the fraction in 2009 was higher.

LHCb

Resource	Site(s)	Used/Pledged
CPU(kHS06*yr)	CERN	52% (*)
	T1	41% (*)
	T2	69% (*)
Disk(TB)	CERN	52% (*)
	T1	28% (*)
	T2	N/A (**)
Tapes (TB)	CERN	27% (*)
	T1	22% (*)

(*) We quote the used / pledged usage, as this is the information given to us by LHCb. Please see below to see the respective percentages of fulfilment of the pledges (which is 80-120%, depending on the resource)

(**) The disk resources in Tier 2s accounts for temporary storage therefore an usage report is not appropriate.

Comments to the LHCb report

In general, LHCb performed the computing tasks according to the plan and did not observe any limitation from the available resources, but note the rather modest quantity of data taken in 2009. The usage of resources is mainly attributed to the MC production and to related analysis tasks.

The usage of CPU at CERN was lower than foreseen due to the lack of data, reducing the amount of analysis and stripping. The usage of Tier2's exceeded the requests and explains somewhat the reduced usage of T1's and CERN for simulation.

Overall, the usage seems to follow the predictions of scaling with running time obtained during the 2009 scrutiny.

Real data taking allowed LHCb to perform some checks of the model. For instance, the data acquisition was done without zero suppression in VELO and data over more bunch crossings was collected. No major problems were observed or are foreseen based on the present experience. Data distribution was also performed according to the plan.

The mDST format was not used in the production. The implementation of this format is prepared now (merging, versioning, and bookkeeping). The Monte Carlo production performed as expected and is decoupled from the data reprocessing activities.

LHCb continues to observe difficulties in running on the grid. The success rate can be as low as 65 % for the first job, then the automatic job submission recover up to 100%. Problems are observed with instabilities in the access to the data, which are mitigated with local copy and CPU 'watch dogs' implemented in the jobs. xrootd will be tested with the hope to improve file

access. A new version of DIRAC installed recently, includes a system to assess the state of the resources. This will allow identifying bad sites and banning them from production. Stability is an issue for both production and analysis (15/17%).

The analysis takes place over GRID and CAF in a transparent way. Non-GRID usage is 30% at CERN. The intensive usage of CERN CAF did not occur in 2009 due to the small data sets. The policy to favor GRID usage will be continued as expected (40% in 2011, 25% after).

LHCb would like to stress again that sharing resources at CERN is a way to optimize the usage of the installed computing power.

ALL EXPERIMENTS

Monte Carlo data production

Simulated data from 2008 or before is not useful anymore and it will be deleted from disk. The experiments have accumulated sizeable samples of simulated data during 2009 in a variety of running conditions: 900 GeV (injection energy), 2.36 TeV, 10 TeV, and 7 TeV. Only the latest one, which the experiments are steadily accumulating at present following the Chamonix meeting, is expected to be useful for physics in the future, and the largest sample at present at 10 TeV centre-of-mass energy will be deleted as the need for disk space will mount.

This less-than-optimal use of the resources is a consequence of the shift in the expected running conditions as the re-commissioning of the LHC progressed and a better understanding of the machine was gained. We feel that there is some room for improvement in the flow of information among the CERN management, the experiments and the reviewing bodies (LHCC and CRSG) that could optimize the experiments' and sites/Tiers planning and prevent repeating similar situations.

Efficiencies

The computing TDR estimates the efficiency to be 85% for CPU and 70% for disk in the case of organized (group driven) analysis, reducing to 60% in the case of chaotic (user-driven analysis). There seems to be some mismatch between these reference numbers and the ones assumed by the experiments.

For simplicity it is assumed that organized activities are by fiat carried out at CERN+ Tier1's while chaotic analysis is carried out only at Tier2's. (Actually CERN is a combination of T0/T1/T2/T3 varying in relative percentage from experiment to experiment) This is not so in the current implementation of the computing models and this is reflected in the fact that the percentage of utilization in the Tier2 even in this early stage is well above the theoretical 60% lying at the 75% when averaged over the Tier2. This results in the Tier2 systematically yielding CPU well above the pledge. This may reflect the fact that the analysis to MC ratio was very small and thus the efficiency will drop in the future. We note that the current efficiency at CERN + Tier1s is 81%, close to the reference value. We recommend that this efficiency factor for the Tier2 be revised.

Efficiency of the utilization of the CPU at Tier 2s per experiment in 2009

ALICE	65%
ATLAS	85%
CMS	66%
LHCb	90%

It is interesting to note that LHCb that actually does only organized activities at the Tier 2 (Monte Carlo production) clearly surpasses the reference value of 85%.

We ask the experimental collaborations and the sites to keep track of the efficiency of the CPU and disk usage to improve the use of the resources, and in order to adjust, as much as possible, the resources available to the needed ones.

Disk usage

While the interpretation of CPU usage is straightforward, disk usage is more subtle to analyse. A metric based exclusively on disk occupancy does not account how frequently the data stored on disk are used. It might well be that disks can be full of data which are rarely accessed and therefore can be stored on tape. We feel that a metric that also takes into account a disk access pattern is more appropriate and also identifies "hot spots", namely data that are frequently accessed. We are aware of the technical difficulties involved and of the timescale needed to define and implement such a metric; for the time being, we require that the disk utilization in the next reports from the experiments should be given in terms of the various data types involved (e.g. RAW, RECO, AOD, derived data, group data, user data and so on) and how frequently they were changed/replaced on disk.

No information has been provided so far on the efficiency of the disk usage so the 70% estimate remains untested.

Sharing of the overall WLCG resources

The following table gives an idea of the degree of utilization of the different experiments of the disk and CPU made available to them through the WLCG. The percentages refer to the fraction of the total mass storage, disk and CPU used per experiment (therefore all columns add up to 100% up to rounding errors). The last column indicates which fraction of the total CPU that a given collaboration has used has been at CERN rather than using the T1's (and, consequently, does not add up to 100%).

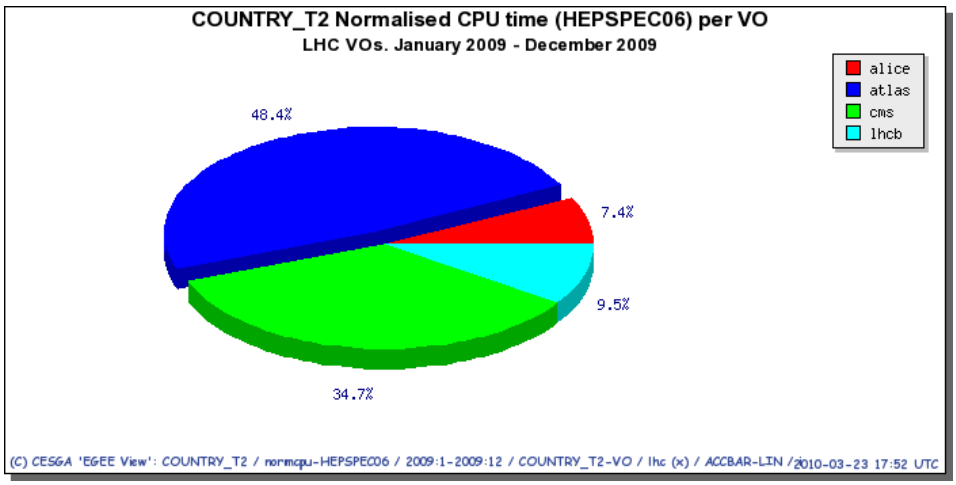
Percentage of use of the resources by experiment in 2009

Collaboration	% of tape in T1+CERN used at end of period	% of disk in T1+CERN used at end of period	% of CPU in T1+CERN used	of which at CERN
ALICE	6%	4%	10%	29%
ATLAS	30%	57%	55%	9%
CMS	61%	34%	26%	35%
LHCb	3%	6%	9%	20%

The differences between the different placing and analysis strategies when it comes to practice are manifest. Note the large usage of CPU by ATLAS in 2009 (more than half of the total available) of the Tier 1 resources. On the contrary ATLAS has been the one that, in proportion, has used less the CERN resources..

CMS has provided a detailed break down of the distribution of their CPU usage: CERN 11%; Tier 1s: 22% and Tier 2s: 67% (installed capacities are 58%, 24% and 18%, respectively). Both CMS and ALICE have heavily relied on CERN for their 2009 work. LHCb is in an intermediate position.

The following figure contains analogous information, but referred to the CPU usage in the Tier 2s. We see that ATLAS accounts for nearly half of the CPU time used.



Delivered versus pledged

The overall level of fulfilment of the pledges can be learned from the following table. The figures refer in all cases to the end of the reporting period (end of 2009)

Resource	Site(s)	Available / pledged
CPU	CERN	120%
	T1	96%
	T2	N/A
Disk	CERN	77%
	T1	83%
	T2	N/A
Tape	CERN	100%
	T1	118%

Some Tier1s have failed to match the expected pledges. From the ATLAS summary we learned for instance that TW-ASGC made available in 2009 only 0.87 M HS06-days of CPU power while the pledge was 2.36 M HS06-days. Other Tier1s have also fallen somewhat short of the pledges.

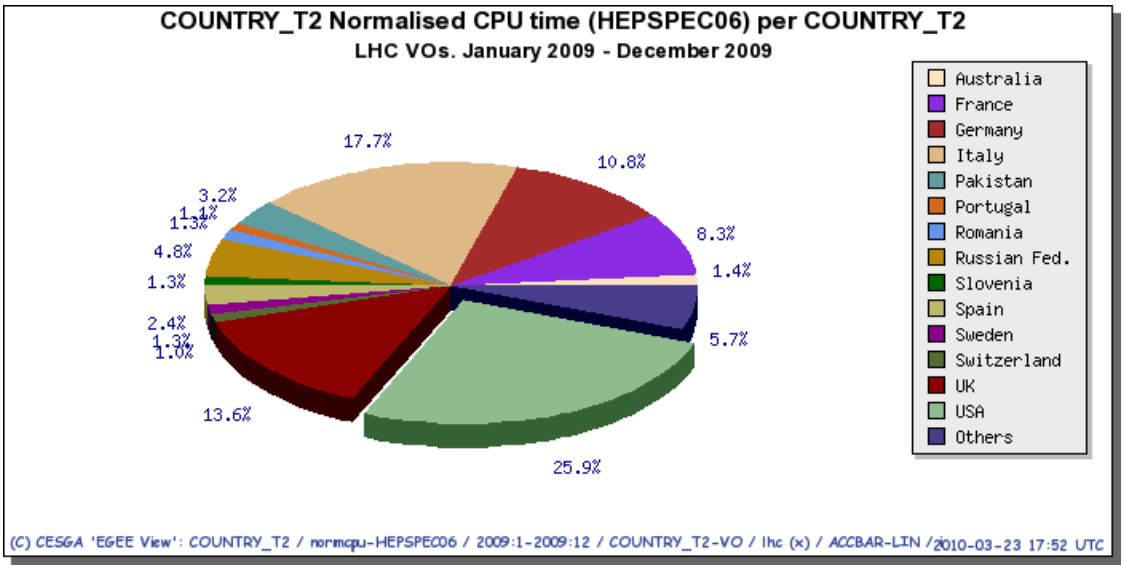
At the end of the reporting period two Tier 1s were below the 90% 'green line' in terms of installed CPU: TW-ASGC and IT-CNAF.

The mismatches are more visible in disk capacity, which is expected to some extent as disk is the most expensive commodity. At the end of the reporting period only four of the sites were above the 90% level, and five were actually below the 80% line: IT-CNAF, BNL, NL-NIKHEF/SARA, CA-TRIUMF, CERN and UK-RAL (listed in order of decreasing mismatch with the pledge).

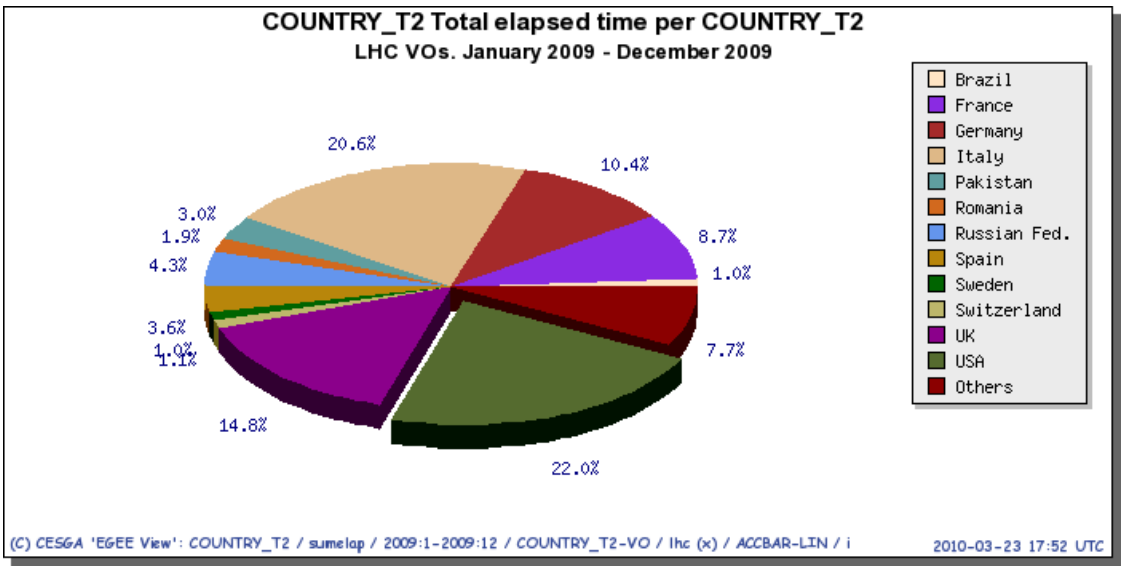
Tape storage was more in line with the pledges generally speaking, but a notable exception was NL-NIKHEF/SARA that had only installed a mere 13% of the pledges. UK-RAL was barely above the 50% 'red line'.

None of these mismatches jeopardized physics analysis or simulated data production in view of the low turnout of the LHC, and to some extent they have been triggered in some cases by the evidence that the agreed pledges were over committing resources. Yet there is a danger in this attitude and it is desirable that in the future the pledged resources are made available within the agreed WLCG deadlines.

The Tier2 situation is extremely irregular. A dual pattern seems to be developing whereby 'good' and 'reliable' sites coexist with sites that fail to fulfil their resource pledges or their reliability/availability targets. The list of sites is obviously too large to refer to individual Tier 2s and we provide here the CPU contributions grouped by country in form of a pie chart.



The following chart displays the elapsed or wall time, from a numerical comparison between the two a measurement of the relative efficiency can be obtained. The overall figures for this have been discussed above.



Centralized reporting from the WLCG is only available for the CPU delivery and therefore only the different experimental collaborations are aware of the disk resource installed, and this perhaps not even in all cases as it has been reported several times in the C-RRB that the flow of information from/to the Tier 2s is not totally satisfactory.

The experiments have provided detailed reports on their usage they have made of the resources made available to them through the WLCG collaboration. The four reports are appended at the end of this document (Part C)..

PART B

2011 scrutiny

Scrutiny of the ALICE request for 2011

Overview

The ALICE computing request was scrutinized by three members of the CRSG. The collaboration provided a first request on March 14th. The referees replied with a series of specific questions by email followed by a phone meeting on March 26th with the ALICE computing representatives. This resulted in a revised version of the ALICE request for 2011/2012 that arrived on March 30th. Unfortunately this did not allow the referees to arrange a subsequent discussion to finalize these recommendations. In the future final requests should be available by March 1st prior to the C-RRB meeting in mid-April.

There have not been significant modifications to the ALICE computing model as a result of the experience gained with data in 2009. However the bulk of the ALICE computing request is based on HI running for which there is still no direct experience. While we gain further insight into the ALICE collaboration's ability to process data using the current model, over the course of the current (2010) year, we expect the first real test to come in early 2011 when they will have recorded and analyzed their first HI dataset.

It is clear that the 2009 ALICE dataset is far too small to really exercise their computing model as they only used about 30% of the computing resources pledged to the experiment last year.

The Current State of ALICE Computing Model

The computing model ALICE is using today is essentially unchanged from the one scrutinized in 2009. The running time and resource efficiency assumptions used by the collaboration are those outlined at the beginning of this document. We note that ALICE event reconstruction requires more than 2 GB/core (the LCG standard) by almost 50%. This could result in inefficient and un-even usage of Tier 1 and Tier 2 resources and could limit the level of resources available to the collaboration when nominal luminosities arrive.

The ALICE collaboration's current CPU requests are based on 50% more CPU/event than assumed in 2009 (46 HE06-s per event now versus 31 HE06-s in 2009 for *pp* events). The HI reconstruction times have grown by a similar factor from 3.8 kHE06-s in 2009 to 5.7 kHE06-s, this year. We also noted that their ESD format has increased in size by nearly one order of magnitude. This may be a fair reflection of the collaboration's current experience but it is clearly a worrying trend. While the growth does not impact the requested resources significantly for *pp* running, the growth for HI data – despite the fact that no HI data has been taken or analyzed – already has an impact. The CRSG would like to understand the source of this growth and request from the ALICE collaboration present a plan to roll it back, at least, to the original model sizes in the not too distant future. We propose to re-visit this in the fall of 2010, when the collaboration will have more experience with 7 TeV collisions and is preparing for the first HI running.

While the ratio of the dataset sizes between *pp* and HI is assumed to be more or less equal – HI events are a factor 12 larger but there are ten times fewer recorded than for *pp*. However it takes one hundred times more CPU to reconstruct each HI representing 90% of the ALICE collaboration's CPU request. Thus their request and eventual performance is subject to much larger uncertainty than for the other experiments that already have experience with their primary (*pp*) datasets.

Scrutiny and Comments

The tables below show the resource estimates for 2011. The results of the present scrutiny are shown in the column "Scrutiny", whereas the ALICE requests are shown in the column "*ALICE"*.

The requested resources are shown in the following table. We should note that in the column labeled 'Scrutiny' we quote the result obtained by our calculations based on the information provided by the computing management. Our scrutiny is based on a very simplified version of their computing model. In the case of CPU and tape usage we agree with their requests to within 10%. We have a harder time scrutinizing their disk needs as much of this is based on the post-reconstruction analysis of the data that is not easy to capture in a simplified model. Still, even in those cases we are able to justify 85% or more of their requests and it is the closest we could come to reproducing their request. Generally speaking, while we have some reservations, we feel that most of the resources requested should be granted in view of the existing uncertainties in this experiment

CERN resources	**2011**	
	Scrutiny	***ALICE***
Tier0 CPU (kHS06)	47	*48*
CAF CPU (kHS06)	12	*14*
Tier0 disk (PB)	5.0	*5.8*
CAF disk (PB)	0.5	*0.5*
CERN tape (PB)	6.7	*6.8*
Non-CERN resources	**2011**	
	Scrutiny	***ALICE***
Tier1 CPU (kHS06)	121	*117*
Tier1 disk (PB)	7.2	*8.5*
Tier1 tape (PB)	13	*13*
Tier2 CPU (kHS06)	111	*121*
Tier2 disk (PB)	6.0	*6.7*

The ALICE collaboration relies most heavily on the CERN resources, as this is where the first past processing takes place and, especially for the HI running, where the bulk of the compression from RAW to ESD (a factor of 10 in event size) takes place. The ALICE collaboration has taken advantage of opportunistic grid computing resources in 2009 – thus boosting their Tier1 usage beyond what is expected in the long term. However, when they are faced with data, especially HI data, they will be even more reliant on CERN resources. We encourage the ALICE collaboration to do what they can to distribute as much of their analysis to LCG sites away from CERN over the long term.

Recommendations

We note that, according to the ALICE request that we received, the storage (both disk and tape) the experiment requires for 2010 and 2011 is covered by the pledges they have received, with the exception of 5% shortfall in disk at CERN (a combination of the T0 and CAF disk resources requested). Since this is well within the uncertainties of our scrutiny we believe ALICE should be in good shape for the first two years of LHC operation.

The same cannot be said for the CPU resources they request. While these fall 10% short at CERN in 2010 (a year for which the scrutiny is already completed) they are projected to be adequate in 2011. Their T1/T2 CPU requests grow substantially in 2011 and the pledges appear to fall more than 20% short before the end of 2011. However we are unable to account for a similar amount of CPU needed from the Tier 1 and Tier 2 centres. If our scrutiny were complete then the current pledges would cover their needs within 10%. Again, this is subject to the remaining uncertainties surrounding processing of the HI datasets.

We believe the ALICE collaboration will be able to analyse the data they will receive in 2010 and 2011 given the LCG computing resources currently pledged. Their request for resources in 2012 will need further study as the collaboration anticipates continued increases will be necessary, at a similar rate to 2011, despite the fact that no new data will be collected by the ALICE experiment in 2012.

Scrutiny of the ATLAS request for 2011

Overview

As mentioned previously, the ATLAS and CMS requests were scrutinized in parallel by the combined team of ATLAS and CMS referees. A meeting with ATLAS representatives was held on March 25th. Further questions were asked and subsequent clarifications were provided by email. The level of detail provided by ATLAS has been excellent and has allowed the referees to reproduce independently the ATLAS computing model to a good level of accuracy. The majority of the requests are firm and justified; some requests in the area of data distribution have a substantial impact on resources and on the long term sustainability of the ATLAS computing model and should be closely monitored.

Generally speaking, the referees feel that the ATLAS computing model is more than adequate to allow efficient data taking and analysis; nevertheless further insight in the areas of data size, processing times, data format and distribution is required in order to match the challenges of steady-state at high luminosity.

ATLAS did not submit a separate request for HI interactions, including the resources needed for this running mode in the pp request. In all cases, the HI requests are within 10% of the ones for pp running.

The ATLAS Computing Model

The model ATLAS is using today and its input parameters are essentially unchanged from those scrutinized in 2009.

Experience with calibration and reconstruction of cosmic ray data has shown that recall speed of raw data from tape may limit ATLAS ability to keep up with the data. As a consequence, the experiment is currently relying much more heavily on disk storage for their reprocessing. This policy is reasonable at the start of the data taking period. However raw data will eventually be kept almost entirely on tape in 2011.

The time needed to generate and reconstruct a simulated event was a point of concern. ATLAS has made progress in order to reduce the required CPU power by 25%. While we welcome the effort which led to this sizeable improvement, we nevertheless insist that the ATLAS collaboration continues these efforts in the future.

The analysis strategy foreseen by ATLAS, organized by trigger streams, implies that analysis is performed on big fractions of the entire data sample; in several cases derived formats are created. In order to make this process sustainable, data are replicated several times in the ATLAS computing clouds so that the associated CPU load can be shared.

We encourage ATLAS to review critically their data distribution model in order to maximize their physics output while keeping the overall resources at a manageable and scalable level. We welcome that ATLAS is planning to take steps in this direction and quote the response given by ATLAS on this issue:

“The ATLAS Computing Model follows closely the recommendations from the Monarch R&D Project and was developed at the time that network bandwidth seemed the scarcest resource. For this reason ATLAS choose to pre-place data for analysis and send the jobs to where the data resides. [….] To cope with the ever-increasing need for disk space a scheme has been developed by which the number of pre-placed copies of least used datasets is decreased. This way to better control the number of copies of our data is currently been implemented and tested and we hope to report on its performance by the end of this year.

Over the past 5 years it has become clear that, contrary to what Monarch concluded, wide area network bandwidth is no longer our scarcest resource but disk space is. Therefore ATLAS could gradually move away from the scheme where jobs always go to the data to a scheme where data could be imported and cached to places where free CPU are available. In such a scheme fewer copies of the data would be needed. ATLAS and CMS have recently approached the CERN IT division to launch an explorative study in this direction. The foreseen time scale for this new data management system is 2012/2013."

Comments and recommendations

The tables below show the resource estimations for 2011. The results of the present scrutiny are shown in the column "Scrutiny", whereas the ATLAS requests are shown in the column "*ATLAS*".

We should note that in the column labeled 'Scrutiny' we quote the result obtained by our own calculations based on the information provided by the computing model (where still applicable) and the information provided in the request document. This number agrees in general rather well with the simplified spreadsheet provided by ATLAS, but is normally below the 'official' ATLAS request, labeled "*ATLAS*". We adhere to the figures obtained from the scrutiny as we have no way of checking the origin of the difference, sometimes noticeable.

Generally speaking, we feel that the resources are well justified, with the exceptions noted below.

The CERN resources are reasonable and should be supported; as noted above, any shortage of resources at CERN would reflect in potential disruption of data, whereas shortage at Tier1s and Tier2s slows down the physics productivity of the experiments.

The recommended resources are shown in the following table. A detailed breakdown is given in the following paragraphs.

CERN resources	**2011**	
	Scrutiny	***ATLAS***
Tier0 CPU (kHS06)	30.2	*30.0*
CAF CPU (kHS06)	45.0	*45.0*
Tier0 disk (PB)	0.9	*0.9*
CAF disk (PB)	5.5	*6.1*
CERN tape (PB)	12.0	*12.2*
Non-CERN resources	**2011**	
	Scrutiny	***ATLAS***
Tier1 CPU (kHS06)	208.1	*226.0*
Tier1 disk (PB)	23.1	*24.8*
Tier1 tape (PB)	30.5	*30.1*
Tier2 CPU (kHS06)	277.5	*278.0*
Tier2 disk (PB)	34.2	*38.4*

Tape resources

Tape request at CERN is consistent with our estimates. The results of the scrutiny for tape requests at Tier1 are consistent with the evaluation shown in the simple model provided by ATLAS.

CERN Tape (PB)	2011	
	Scrutiny	*ATLAS*
Pre-existing data	8.86	*8.90*
LHC data	3.17	*3.30*
Total	**12.04**	***12.20***

Tier1 Tape (PB)	2011	
	Scrutiny	*ATLAS*
LHC RAW data	4.39	*4.60*
Real ESD+AOD+DPD data	9.00	*7.20*
Simulated data	12.29	*13.80*
Group+user	0.20	*0.20*
Pre-LHC data	4.10	*4.30*
Cosmics	0.50	
Total	**30.49**	***30.10***

Disk resources

There is a substantial increase in the CAF disk request as compared to the 2010 resources and also to the original projection of the computing model. This is partly due to storing new and old versions of all AOD and DPD real and simulated data on disk in order to allow a more rapid software validation, a policy that will not be sustainable in the future, but acceptable in the initial period. Our itemized scrutiny is in slight discrepancy with the request, although the overall figures are in better agreement.

T0 Disk (PB)	2011	
	Scrutiny	*ATLAS*
Buffer for RAW & processed data	0.58	*0.65*
Buffers for I/O	0.30	*0.20*
Total	**0.88**	***0.85***

CAF Disk (PB)	2011	
	Scrutiny	*ATLAS*
Fraction of current RAW data	0.63	
Real AOD+DPD data	2.37	*3.90*
Simulated RAW+AOD+DPD data	1.54	*1.20*
Calibration and alignment outputs	0.30	
Group data	0.30	*0.60*
User data (scratch)	0.40	*0.40*
Total	**5.54**	***6.10***

The Tier1 disk requirement is dominated by the storage of real data, as well as simulated data. In both cases, most of the disk space is needed for ESD data. Non-negligible disk space is needed for processing and I/O buffers (about 0.3PB per Tier1), and to store data skimmed and slimmed by physics/detector groups. The latter has been computed by assuming that it corresponds to 20% of the total space required to store AOD data. We believe this assumption should be justified based on real experience with data; the disk space reserved for group data can be eventually minimized by fully exploiting the derived physics data format.

When computing the disk space needed by real and simulated data, we take into account the fact that ATLAS, contrary to pp collision data, does not plan to store real and simulated ESD data from HI collisions on Tier1 disk.

Tier1 Disk (PB)	**2011**	
	Scrutiny	***ATLAS***
Current RAW data	0.78	*0.70*
Real ESD+AOD+DPD data	9.45	*11.00*
Simulated RAW+ESD+APD+DPD	5.70	*6.10*
Calibration and alignment outputs	1.00	*1.00*
Group data	2.45	*2.30*
User data (scratch)	0.60	*0.60*
Cosmics	0.20	*0.20*
Processing and I/O buffers	2.90	*2.90*
Total	**23.08**	***24.80***

The disk space at Tier2 is also dominated by real and simulated AOD+DPD data. The same comments made about group space at Tier1 apply to Tier2 as well.

When computing the disk space needed by real collision data, we take into account the fact that ATLAS, contrary to pp collision data, does not plan to store derived data (dESD) from HI on Tier2 disk. Furthermore, we assume that only 2 copies of AOD from HI collisions are kept on Tier2 disk, instead of 10 copies as in pp collision data.

Tier2 Disk (PB)	**2011**	
	Scrutiny	***ATLAS***
Current RAW data	0.26	*0.30*
Real AOD+DPD data	16.42	*19.70*
Simulated RAW+AOD+DPD data	12.50	*11.10*
Calibration and alignment outputs	0.30	*0.30*
Group data	2.29	*4.60*
User data (scratch)	1.80	*1.80*
Processing buffers	0.60	*0.60*
Total	**34.17**	***38.40***

We would like to emphasize, as in previous reports, that the AOD + DPD file distribution policy in the Tier 2 s of ATLAS is quite ambitious. We understand that it is a priority of ATLAS to stick to this policy of wide distribution, but we have doubts as to the sustainability of this in the medium term as the volume of the data increases. A distribution policy based on earlier data streamlining and fewer copies (4 or 5, rather than 10) would represent substantial savings without compromising the effectiveness of the model. We strongly welcome the work plan being discussed within ATLAS in order to address this issue in the medium term.

CPU resources

The Tier0 requests are firm and should be granted. The CAF requests are essential to monitor the detector physics performance in quasi-real time. While the requests for partial reconstruction and non-automatic calibration can be computed from the equivalent tasks performed at the Tier0, it was not possible to scrutinize from first principles the "group activities", which are related to detector performance studies and offline checks. The CPU request for servers might be reduced by profiting of virtualization.

Tier0 CPU (kHS06)	**2011**	
	Scrutiny	***ATLAS***
Full reconstruction	16.0	*16.0*
Partial processing and valdation	3.2	*3.0*
Merging and monitoring	4.0	*4.0*
Automatic calibration	7.0	*7.0*
Total	**30.2**	***30.0***

CAF CPU (kHS06)	**2011**	
	Scrutiny	***ATLAS***
Partial reco, debug, monitoring	5.0	*5.0*
Non-automatic calibrations	5.0	*5.0*
Group activities	19.0	*19.0*
User activities	4.0	*4.0*
Servers	12.0	*12.0*
Total	**45.0**	***45.0***

The Tier1 CPU requests are now dominated by Monte Carlo production. As a matter of fact, ATLAS changed the share of simulation production between Tier1 and Tier2 from 40:60 to 60:40. Therefore, the load on Tier1 CPU has now increased. A non-negligible amount of computing power is required by the reconstruction and merging of simulated events, which takes more CPU power than what required by real data. The contribution due to group and user activities correspond to reasonable assumptions on the number of physics groups, on how often they pass on the entire datasets and on the CPU power required to process a single event. We note that, contrary to CMS, data skimming and slimming are not done centrally and

in parallel to data reprocessing, but at the group level and more frequently. The corresponding resources are therefore higher by a factor of about 10. We encourage ATLAS to explore the possibility of and eventually implement skimming and slimming in a more centralized way.

ATLAS requests an additional 10% (5%) to the computing power needed to reprocess (simulate) events, as a contingency to take into account CPU power spent for validation studies at the beginning of each reconstruction or simulation cycle. We feel this is an overestimate.

There is a discrepancy between the simplified spreadsheet provided to us and the estimate declared in the ATLAS request in what concerns the CPU time needed to reconstruct and merge MC events ("Simulation reconstruction" in the table below). We adhere to the time stated in the request.

In our estimates, we take into account the different reconstruction and simulation times needed for HI collision events with respect to pp collision events.

Tier1 CPU (kHS06)	**2011**	
	Scrutiny	***ATLAS***
Re-processing	31.1	*34.0*
Simulation production	98.0	*109.0*
Simulation reconstruction	25.9	*34.0*
Group (+user) activities	53.0	*49.0*
Total	**208.1**	***226.0***

The Tier2 CPU requirements are mostly due to user analysis activities and, to a lesser extent, to group activities and simulation production. The group and user activities have been computed by using reasonable assumptions on processing times, number of passes and time required for a single pass on the datasets.

Tier2 CPU (kHS06)	**2011**	
	Scrutiny	***ATLAS***
Simulation production	65.4	*65.0*
Group activities	53.0	*49.0*
User activities	159.1	*164.0*
Total	**277.5**	***278.0***

Scrutiny of the CMS request for 2011

Overview

The ATLAS and CMS requests were scrutinized in parallel by a single team of referees. After the preliminary joint review together with the LHCC on February 16th, a formal request of resources for pp collisions was received in the first days of March. A few days later a request for resources specific to the HI runs was received. This is scrutinized also separately at the end of this report.

A phone conference meeting among the CMS representatives and the CRSG referees took place on March 26th, with a sheet of questions/answers being exchanged. We also had several email contacts with the computing management representatives of CMS.

The level of information provided by CMS has been very good and allowed the referees to reproduce independently the CMS computing model to a good level of accuracy. Our calculations agree well with a simplified spreadsheet provided by CMS. We find however a systematic bias of the formal request with respect to our scrutiny without additional information being provided about this. The numbers quoted below under the 'Scrutiny' column correspond to the result of our calculations that, as mentioned agrees with the estimates of the simplified spreadsheet of CMS except for a few points of discrepancy that we shall review below.

The interaction with the CMS representatives has been extremely fluid and we are grateful to them for their ample collaboration.

The majority of the requests is firm and justified, given the special data taking scenario for 2010.

Generally speaking, the referees feel that the CMS computing model is adequate to survive the coming data taking period in 2010; some minor changes might be required in order to match the challenges of 2011. Further insight in the areas of data size, processing times, data format and distribution will probably be required in order to match the challenges of steady-state at high luminosity.

The CMS computing model

The CMS computing model has not undergone significant conceptual changes since it was first published in 2005. Important parameters such as the data-taking rate, RAW and RECO sizes and the CPU required to process a single proton collision have not changed since last year's review.

However the AOD grew substantially in 2009 and no changes are expected on this until 2012. This is partly compensated by a reduction in the size of RAW MC in 2010 and of RAW in 2011.

The collision overlap in the 2011 run will increase; as the LHC expects to deliver the same current in fewer bunches. Simulations studies carried out by CMS show that this will require more CPU per event for reconstruction. But the number of overlaps appears to have been reduced with respect to previous estimates, possibly in view of the relative low luminosity that the LHC is expected to deliver. Reconstruction time is in any case assumed to go up by a 25% in 2011 with respect to 2010.

CMS now plans to reconstruct data at the Tier 0 in 2010 and 2011 at a 75% of the nominal incoming rate (300Hz), as opposed to the 50% assumed until now. This increases the need for resources at the Tier 0, but it seems acceptable to the CRSG as this is in the critical path.

CMS plans for the equivalent of two full re-processings during 2011, but the assumed time for reprocessing increases as the total volume of data collected to 60 days in 2010 and 90 days in 2011.

CMS plans to stream events based on trigger information (electrons, muons, jets, missing E, etc.) If an event fires both the electron and jet trigger (something their simulation says should happen 20% of the time in 2009) then it will be streamed to both datasets and thus will be reconstructed (and re-reconstructed) twice. A careful analysis of their trigger algorithms leads them to conclude that there will be a 40% increase in the number of events on analysis streams, relative to the number of raw triggers, resulting in a 40% increase in the Tier0 and Tier1 event processing CPU and a similar increase in the storage needs for RAW and RECO datasets. Based on their simulation studies and the validation of these studies using the 2009 datasets CMS anticipates reducing the trigger overlap in 2011 to 25% (previously it was 20%). We believe that CMS should work harder in trying to reduce the data set overlap as much as possible.

We note that due to the larger data acquisition rate and the large overlap fraction the total volume of data accumulated by CMS and stored on tape nearly doubles that of ATLAS. However, the final disk requests for CMS are comparable to ATLAS's at Tier1 and about half at Tier2, due to the different data distribution strategies between the two experiments.

Requests and recommendations

The tables below show the resource estimations for 2011. For each year, the results of the present scrutiny are shown in the column "Scrutiny", whereas the CMS requests are shown in the column "CMS".

Generally speaking, we feel that the resources are well justified, with the few exceptions noted below. The CERN resources are reasonable and should be supported; as noted above, any shortage of resources at CERN would reflect in potential disruption of data, whereas shortage at Tier1s and Tier2s slows down the physics productivity of the experiments.

The recommended resources are shown in the following table. A detailed breakdown is given in the following paragraphs.

CERN resources	**2011**	
	Scrutiny	***CMS***
Tier0 CPU (kHS06)	71.9	*72.0*
CAF CPU (kHS06)	34.1	*34.1*
Tier0 disk (PB)	0.9	*0.9*
CAF disk (PB)	3.6	*3.6*
CERN tape (PB)	21.6	*21.6*
Non-CERN resources	**2011**	
	Scrutiny	***CMS***
Tier1 CPU (kHS06)	139.5	*150.7*
Tier1 disk (PB)	18.0	19.5
Tier1 tape (PB)	49.4	52.4
Tier2 CPU (kHS06)	305.1	*319.5*
Tier2 disk (PB)	18.1	*19.9*

Tape requests

An essential step of data taking is the validation of binary streamers from the CMS DAQ before discarding them and saving the corresponding ROOT I/O files. CMS used to save binary streamers to tape and recycle the tapes after validation. They now plan to use a disk buffer for that. This is a more technically viable solution to provide space during the validation than writing many small files to tape and then recycling the tapes. We therefore welcome this solution and accept their plan to store back at CERN data which will be reconstructed in external Tier1s for redundancy.

We find that the tape requests at Tier1 generally justified.

CERN Tape (PB)	2011	
	Scrutiny	*CMS*
Group data (CAF)	3.60	*3.60*
LHC and preexisting data	17.98	*18.00*
Total	**21.58**	***21.60***

Tier1 Tape (PB)	2011	
	Scrutiny	*CMS*
LHC RAW data	6.51	*6.77*
Real ESD+AOD data	13.89	*14.63*
Simulated data	24.95	*27.05*
Group+user	1.00	*4.00*
Pre-existing data	3.00	
Total	**49.35**	***52.45***

Disk requests

The CMS disk request at Tier0 is reasonable and justified as the Tier0 is on the critical path

T0 Disk (PB)	2011	
	Scrutiny	*CMS*
Buffer for RAW&processed data	0.61	*0.61*
Buffers for I/O	0.25	*0.24*
Total	**0.86**	***0.85***

Rather than requesting for given amounts of RAW, RECO, AOD and simulated data, CMS organizes CAF disk according to activities (“Physics”, “Commissioning”, “Alca”, etc.), but they not give any details on each single request. The request of 960 TB in 2011 for commissioning, embedded in the requests for real and simulated data in the table below, seems unnecessarily

large to us. With this exception, given the critical service provided by the CAF, we warrant the resources requested by CMS.

CAF Disk (PB)	2011	
	Scrutiny	***CMS***
Fraction of current RAW data	0.00	*2.16*
Real AOD data	0.80	
Simulated RAW+AOD data	1.46	
Calibration and alignment outputs	0.43	*0.67*
Group data	0.40	*0.40*
User data (scratch)	0.40	*0.40*
Total	**3.61**	***3.63***

The **Tier1 disk** requests are obtained as follows:

- For **real data**, there are on Tier1 disk: 20% of the overall RAW data; a full copy of the current RECO version and 10% of each of the previous two RECO versions over the whole Tier1 ensemble; 7 copies of the current AOD version (1 copy per Tier1), no copies of previous AOD versions
- For **simulation**, in both 2009 and 2010, there are on Tier1 disk: 10% of the overall sRAW data, 10% of the current sRECO version over the whole Tier1 ensemble; 7 copies of the current sAOD version (1 copy per Tier1), no copies of previous versions

The disk space is dominated by RECO data (4PB for real and simulated events) and by the real and simulated AOD (10.7PB); the latter is due to the factor 7 replica for AODs.

Tier1 Disk (PB)	2011	
	Scrutiny	***CMS***
RAW data	1.53	*1.31*
RECO data	3.51	*3.69*
sRAW data	1.14	*1.47*
sRECO data	0.46	
AOD (real+simu)	10.72	*13.01*
Analysis space	0.00	*0.00*
WAN buffers	0.60	*0.00*
Total	**17.97**	***19.48***

The **Tier2 disk** requests are obtained by assuming that the Tier2 ensemble contains: 4 global copies of 90% of the current AOD (real and simulated) version; 4 global copies of 10% of the current RECO (real and MC) version.

The rest of the Tier2 disk space is used for group data and MC processing buffers
The analysis activities are adequate. The group data has been obtained by multiplying the number of Tier2 (40) by the space allocated to groups and users in each Tier2 (80TB), giving 3.2PB.

Tier2 Disk (PB)	2011	
	Scrutiny	*CMS*
Real RECO	1.17	*1.47*
Real AOD data	4.22	*5.30*
Simulated RECO+AOD data	8.68	*9.14*
Calibration and alignment outputs	0.00	*0.00*
Group data	3.20	*3.20*
Processing buffers	0.86	*0.80*
Total	**18.13**	***19.91***

CPU requests

The Tier0 CPU resources needed for full reconstruction are straightforward to compute starting from the trigger rate end event processing time. CMS plans to increase this to 225 Hz, i. e. 75% of their trigger rate in 2010 and 2011

Since these are on the critical path, the resources required for these tasks should be warranted.

Tier0 CPU (kHS06)	2011	
	Scrutiny	*CMS*
Prompt RECO	46.9	*47.0*
Express stream and repack	9.0	*9.0*
VO servers, validation	15.0	*15.0*
Reconstruction of calibration stream	1.0	*1.0*
Total	**71.9**	***72.0***

The **CAF CPU** requirements have been redefined; commissioning, calibration, alignment and physics monitoring activities, which played a dominant role in last year's requests, now account for nearly half of the request, the other half being due to "CERN analysis resources". The latter account for users and groups performing analysis activities and being geographically associated to CERN. The resources requested amount to about 10% of the corresponding resources requested for Tier2s. This means that the CERN CAF is considered as a Tier2 facility of nearly four times bigger than the average CMS Tier2.

We feel compelled to grant this request, but we ask the experiment to adhere to the approved computing model and make a balanced usage of CERN/ nonCERN resources in order not to pre-empt the GRID philosophy.

It should be stated nevertheless that these resources are comparable to those of ATLAS. ATLAS however has overall a larger volume of resources.

CAF CPU (kHS06)	2011	
	Scrutiny	*CMS*
Calibration and alignment	8.0	*8.0*
CERN analysis resources	16.5	*16.5*
Low latency analysis	8.5	*8.5*
Interactive usage	1.1	*1.1*
Total	**34.1**	***34.1***

The **Tier1 CPU** requests are dominated by reprocessing of real and simulated data, each contributing half of the total. A small amount of CPU is needed for skimming. This used to be about a factor 10 bigger in the past; the reason for this reduction is that CMS now plans to perform skimming as a centralized activity performed in parallel with reprocessing, instead of a group-based and frequent one.

Contrary to ATLAS, CMS does not expect to run large scale simulations during reprocessing at the Tier1. For this reason, the CMS resources at the Tier1 are significantly smaller than ATLAS's.

Tier1 CPU (kHS06)	2011	
	Scrutiny	*CMS*
Re-processing of real data	70.2	*146.4*
Re-processing of simulated data	65.6	
Group activities (skimming)	3.8	*4.3*
Total	**139.5**	***150.7***

The Tier2 CPU requests are reasonable. Assumptions similar to the ones made for ATLAS lead to similar results.

Tier2 CPU (kHS06)	2011	
	Scrutiny	*CMS*
Simulation production	115.1	*120.0*
User activities	190.0	*199.5*
Total	**305.1**	***319.5***

Heavy Ion request

For the first time CMS requests separately resources for the heavy ion run in during 2011. The request was contained in a detailed document. The CRSG understands that there are large unknowns in the estimates and that the real needs can be fully assessed only after some experience.

The CMS request for HI running is summarized in the table below

Resource	2011
Tier 0 CPU	*No additional request*
Tier 0 Disk	*No additional request*
Tier 0 Tape (PB)	***2***
Reconstruction CPU (kHS06) (Tier 1 services)	***5.8***
Reconstruction and data serving disk (TB) (Tier 1 services)	***875***
Tier 1 Tape (TB)	***800***
Tier 2 CPU (kHS06)	***9***
Tier 2 Disk (TB)	***500***

The reconstruction CPU and data serving disk will be provided by Vanderbilt University, while Tier 1 tape will be provided by FNAL.

The CRSG endorses the above request.

Scrutiny of the LHCb request for 2011

The LHCb computing resources in 2011 are estimated using a model very similar to the one discussed and approved in the last scrutiny exercise. The experience based on data taking in 2009 was useful to test some assumptions (event size, trigger rates) but did not lead to any fundamental change in the computing model.

This CRSG referees have met the LHCb computing team twice (on February 22nd and on April 7th) and have a fruitful email exchange. A simplified model was made available to the referees and reproduces the global features of the LHCb computing model and is in agreement with the resources request. We would like to thank the LHCb colleagues for the collaborative attitude and for the detailed answers. We note however a large delay in providing the official requests with respect to the schedule agreed last year and we would like to stress that the deadline for requests to RRB should be scrupulously followed in the future.

LHCb running scenario

The running time estimates used by LHCb are based on the agreed machine running scenarios. They are summarized in Table 1 , with some adjacent assumption on availability and efficiency.

LHCb will not take data during the heavy ion runs.

Table 1: LHC running schedule 2010-2011.

Period		Running Mode	Availability	Efficiency	All	
					running	Total
1-mars-10	1-avr-10	p-p	0.3	0.3	2.41E+05	
1-avr-10	16-oct-10	p-p	0.7	0.4	4.79E+06	
16-oct-10	1-nov-10	HI	0.7	0.4	3.87E+05	
1-nov-10	1-déc-10	HI	0.7	0.4	7.26E+05	1.1E+06
1-déc-10	1-janv-11	Shutdown	0	0	0.00E+00	
1-janv-11	1-févr-11		0	0	0.00E+00	
1-févr-11	1-mars-11	p-p	0.4	0.3	2.90E+05	
1-mars-11	1-avr-11	p-p	0.7	0.4	7.50E+05	6.1E+06
1-avr-11	16-oct-11	p-p	0.7	0.4	5.18E+06	5.2E+06
16-oct-11	1-nov-11	HI	0.7	0.4	3.87E+05	
1-nov-11	1-déc-11	HI	0.7	0.4	7.26E+05	1.1E+06
1-déc-11	1-janv-12	Shutdown	0	0	0.00E+00	

The period for 2010 is defined as the period between the start of LHC in 2010 and April 1st 2011 (when the 2011 pledges have to be made available) while the running time for 2011 is assumed to encompass the period from April 2011 to mid October 2011. No data collection is expected at LHCb during the heavy ions running.

LHCb request for 2011

The request of computing resources is summarized in Table 2.

Table 2: LHCb resources requests for 2011

Date	Site	kHS06	Disk (TB)	Tape (TB)
April 2011	CERN	21	1510	2480
	Tier-1	65	3500	3470
	Tier-2	36	20	0

The CPU power is the parameter promoted by LHCb to represent the CPU needs, as shown in Table 2. The available power is important for the reprocessing steps and also for the MC production campaigns that may focus in a short period of time resources beyond the normal usage. Nevertheless, an efficient usage of resources is important and in case of large latencies observed during the normal running conditions, changes in the functional model and in the interactions with the computing centers should be proposed. The integrated CPU power was also presented as an alternative, in order to compare with other experiments, and is presented in Table 3.

Table 3: Integrated CPU requests.

kHEP06*year	2011
(CERN T0)	3,07
(CERN CAF - Analysis/Calib/Align)	11,91
Online Farm	
CERN T0 + T1	17,26
Tier1s	48,91
Tier2s	31,48
Total	97,65

The distribution of the computing power among sites follows the pattern already presented last year. The observation during 2009 is slightly different mainly due to lack of data and to an excess in Monte Carlo production. However the distribution is expected to follow the model in presence of significant data collection.

Following the recommendation of the last year scrutiny, the disk resources include the mass storage system (MSS) front-end disk cache – about 70 TB per site. This resource is considered as an upper estimate and is managed by the sites in order to optimize the performance of the MSS.

The main difference between 2011 and 2012 requests is related to the difference in running time, from 6.1 10^6 to 5.2 10^6 and the migration of analysis outside CERN, expected to reach 25% at CERN in 2011 from 40% in 2010.

The strategy to use the on-line farm during the shutdown cannot be applied to 2010/2011 shutdown which is expected to be short and will be also used for the maintenance of the cooling system. In order to compensate for this change, work is progressing on re-accessing the data from CASTOR in order to perform the reprocessing. The usage of the farm will remain possible for MC production during the short period of time remaining in the shutdowns after the cooling operations. In 2012 the farm usage should be possible but it is not included so far in the resource model.

The number of reprocessing passes is assumed to be 3 in 2010 and 2 in 2011. These passes consider only the data of the respective years and it is expected that a full data reprocessing will take place in 2012. The processings will be accompagned by a stripping, however the stripping is in principle decoupled and can be performed as soon as a significant understanding of the data requires a re-made of the selected samples. The numbers of strippings are assumed to be 4 in 2010 and 2011.

The toy Monte Carlo processing is considered as an analysis related activity and is included in the analysis resources. This follow therefore the strategy of the decrease of analysis activity at CERN in 2011. The split between Toy MC jobs -with minimal data access - and the real analysis with intensive data access will have to be re-assessed at the end of 2010 in light of first data taking.

A new, reduced data format microDST was proposed last year in order to speed-up the analysis. The format is at present in commissioning with the physics working groups. It is planned to be managed centrally, taking advantage of the production mechanisms that allow a coherent strategy (central production, bookkeeping etc.). The associated resources are absorbed at present in the analysis requests.

It should be noted that the trigger rate is assumed to be 2KHz in the model, but not yet reached during the first (assumed) month of data taking (rates around 100 Hz only have been reached and the startup is slow). The difference may be considered as a contingency and the trigger strategy is obviously tightly dependent on the increase in instantaneous luminosity at LHC. The difference observed in 2009 between the requests and pledges (see 2009 usage report) should be closely followed in the next period.

Conclusions

The resources requests for 2011 are based on the realistic estimates already thoroughly discussed during the previous scrutiny. Some assumptions related to the data taking have been tested, albeit with modest data set taken in 2009. Some other assumptions are still to be confronted to an intensive data taking, as expected for this year.

We conclude that the requested resources (Table 2) are realistic and will fulfill the needs of the LHCb experiment for the 2011 run period. Given the rapid change of the computing intensity during 2010, we would like to recommend an intermediate scrutiny by autumn 2010 in order to prepare the resources scrutiny round of 2011.

Comments on the experiments forecast for 2012

The experiments have submitted to the CRSG their expected request for 2012 as well, on the assumption that the running conditions will not change with respect to the expected ones at present.

The profile of the expected requests in 2012 is flat with a slight increase (of the order of 10% or less) in some of the resources.

The CRSG will report on these requests in the October 2010 C-RRB meeting, ahead of the 2011 C-RRB so that the funding agencies, tiers and sites can draw their procurement plans

PART C

Experiments usage reports

ALICE
Yves Schutz
Friday, March 26, 2010

COMPUTING RESOURCES USAGE IN 2009

The usage of computing resources by ALICE in 2009 is briefly discussed. Detailed data on the usage of CPU cycles and disk storage are outlined in the WLCG monthly accounting reports for Tier 0 and Tier 1s[1] and for Tier 2s[2]. The ALICE accounting report, including the non-LCG sites, can be retrieved from the ALICE MonALISA web pages[3].

Data collection

In 2009, ALICE has recorded cosmic data and calibration starting in June 2009 and pp collision data at 900 GeV and 2.36 TeV during November 2010. Data were taken with various configuration of the magnetic field and of the trigger:

- A total of 0.988×10^6 pp collision events for a total raw data size of 371 GB have been registered.
- A total of 1.15×10^9 cosmic events for a total raw data size of 85 TB have been registered.
- The total amount of calibration data amounts to 36 TB.

All the raw data have been stored in the CASTOR mass storage system at CERN. Only collision data have been replicated 2 times (all together 3 copies) in Tier 1's mass storage.

Data processing

All raw cosmic and collisions data have been reconstructed once at the Tier 0 and several times at the Tier 1s. So far, the collision data have been reconstructed 6 times with continuously improved calibration and alignment parameters. The success rate of the reconstruction is 99%. The size of the ESD is 0.290 GB/event, a factor 10 larger than the value used in the computing model. The ESDs have been replicated 3 times and distributed in the Tier1s and Tier 2s xrootd-enabled storage elements.

The ESDS have been analyzed through analysis trains with different wagons, provided by the physics working groups, and through end-user analysis on the GRID. Trains are run several times over the entire set of ESDs. So far 15 trains have been run. Analysis is run indifferently on Tier 1s and Tier 2s where ESDs are stored.

During data taking, a fraction of the events of each reconstructible run are reconstructed on the CAF and the resulting ESDs are analyzed to provide basic QA information.

---

[1] http://lcg.web.cern.ch/LCG/accounting/Tier1/2009/december-09/Master_accounting_summaries_December2009.pdf

[2] http://lcg.web.cern.ch/LCG/accounting/Tier-2/2009/december-09/Tier2_Accounting_Report_December2009.pdf

[3] http://pcalimonitor.cern.ch/reports/

Monte-Carlo production

In 2009 until today, 70 Monte-Carlo production have been run. Most of the productions are intended for the evaluation of efficiencies and systematic errors needed by the first physics analysis. Productions for 2 energies, with various magnetic field settings, different generators, and running condition data have been performed. Other productions involve rare signals for related physics performance studies. One set of AA data was produced with and without quenching. ESDs, Monte-Carlo truth data, and detectors debug data have been saved and replicated 3 times in the xrootd enabled storage elements.

Obsolete data are continuously deleted.

Accounting

The resources used by ALICE in 2009 are summarized in Tab 1. Data have been collected from the WLCG accounting reports and from the ALICE MonALISA monitoring.

Tab 1: *Usage of resources in 2009. CPU and Disk are average over 12 months, tape is the amount used by December 2009. T0 and T1 data are taken from the WLCG accounting and T2 data from the MonALISA monitoring.*

	T0+CAF	**T1**	**T2**
CPU (KSI2K)	1826	3000	4414
Disk (TB)	343	532	365
Tape (TB)	1224	620	0

The profile of running jobs in the past year is displayed in Fig. 1 and Fig. 2.

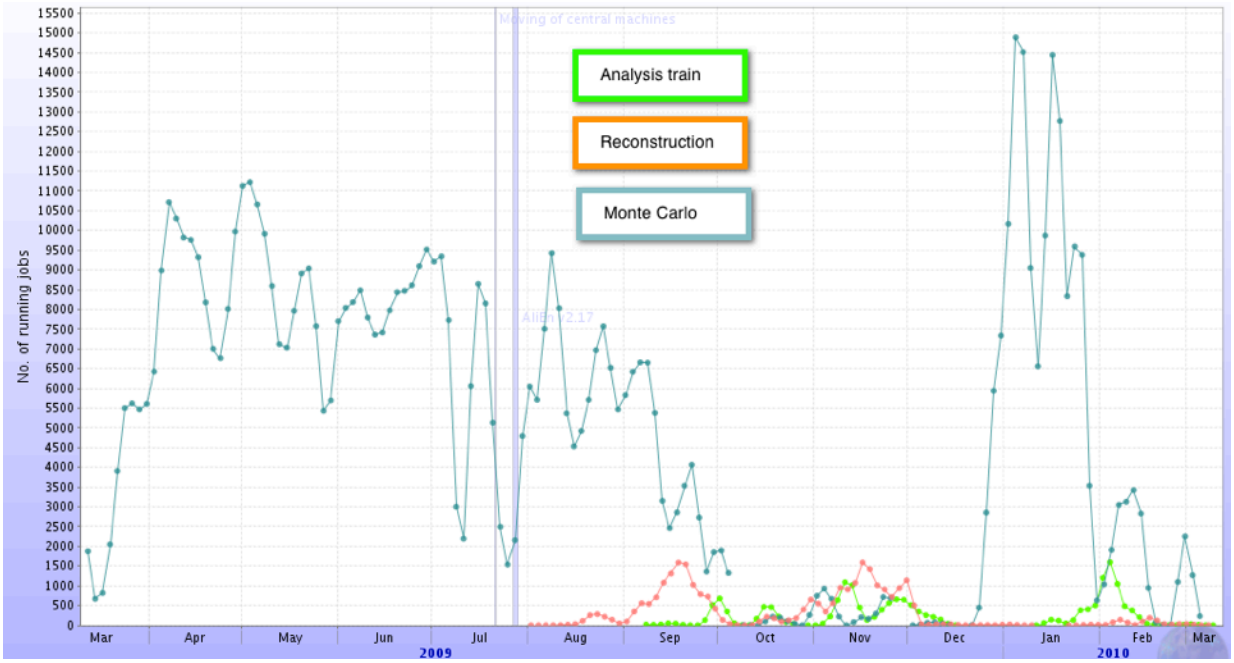


Fig. 1: *Past year profile of the number of jobs running for Monte-Carlo production, cosmic and collision data reconstruction and for the analysis train.*

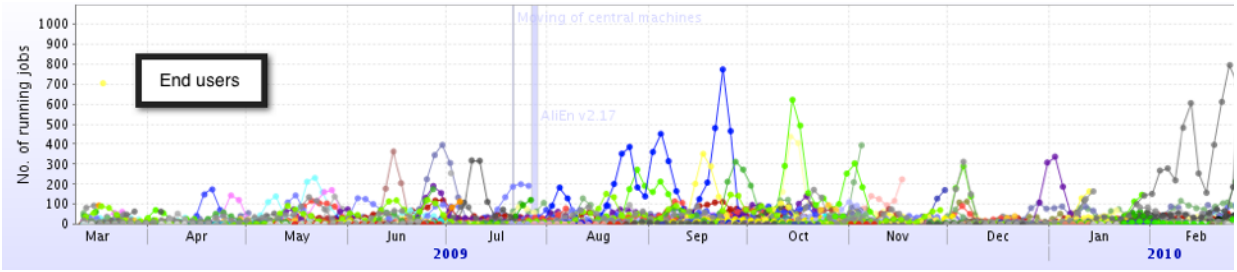


Fig. 2: *Past year profile of the number of jobs running for end-user analysis. About 200 users submit regularly jobs.*

Usage of ATLAS computing resources in 2009

The ATLAS Collaboration*

8 March 2010

1. Introduction

This report summarizes the computing resources used in the calendar year 2009 by the ATLAS Collaboration at CERN and elsewhere. Computing resources are made available to ATLAS through the WLCG (World-wide LHC Computing Grid) infrastructure.

WLCG provides also a number of monitoring, reporting and accounting tools. Most of the information in this report is extracted from reports provided by the computing centres themselves, using these tools.

Main ATLAS computing activities can be grouped into a small number of categories:

- Data management (distribution, deletion, etc.)
- Tier-0/CAF processing of LHC data from the ATLAS detector in real time
- Reprocessing of real data on the Grid
- Production of simulated data on the Grid
- User analysis on the Grid

All these activities use middleware tools in one of the three Grid flavours supported by ATLAS (EGEE/gLite, OSG/VDT, NorduGrid/ARC), as well as the software layers developed by ATLAS to interface the users to the middleware and provide additional functionality:

- DDM (Distributed Data Management) uses the LFC file catalogues and the FTS data transfer system, and the SRM interfaces to different kinds of local Storage Elements
- T0MS (Tier-0 Management System) uses the Castor data storage at CERN and the LSF job scheduler
- ProdSys/PanDA distributes production jobs to the Computing Elements of Grid sites
- GanGA/pAthena distribute analysis jobs to the Computing Elements of Grid sites and provide the user access to Grid computing resources

2. Global accounting at CERN, Tier-1 and Tier-2 sites

Figures 1a, 1b and 1c provide a global view of the usage of CPU and disk resources at CERN and Tier-1 sites. The histograms show the ATLAS usage as reported by the sites through WLCG [1]. The dashed lines show the pledged capacities for 2008 (valid through September 2009) and 2009 (valid from October 2009), scaled by the standard efficiency factors (85% for CPU and 70% for disk) [2]. The solid lines show the actually installed

* Editor: D. Barberis

capacities, also scaled by the standard efficiency factors. The CPU line is also scaled by the reported average availability for ATLAS each month [3].

Figure 1a: CPU and disk space used at CERN and some Tier-1 centres in 2009 (histograms), compared to resource pledges (dashed pink lines) and installed capacities for ATLAS (solid green lines).

Figure 1b: CPU and disk space used at some Tier-1 centres in 2009 (histograms), compared to resource pledges (dashed pink lines) and installed capacities for ATLAS (solid green lines)

Figure 1c: CPU and disk space used at some Tier-1 centres in 2009 (histograms), compared to resource pledges (dashed pink lines) and installed capacities for ATLAS (solid green lines)

It is evident from the figures that at most sites we have used close to 100% of the available resources, concerning CPU and disk space. CPU usage has in some sites and during some periods exceeded the ATLAS allocation, as idle resources at ATLAS sites can be used opportunistically if the site allows it. Disk servers that came online at the end of 2009 are meant for the forthcoming 2010 data and will not be filled until enough new data has been collected; indeed, part of the existing data on disk will have to be removed to make room for new real and simulated events.

Figure 1d shows the cumulative usage statistics for all ATLAS Tier-1 centres. Figure 1d (right) shows that the installed disk space has been always filled within 3-4 months, and once the space taken by the buffers is taken into account, we had almost all disk allocated to real and simulated data full throughout the year. See Section 5 for more details of disk usage.

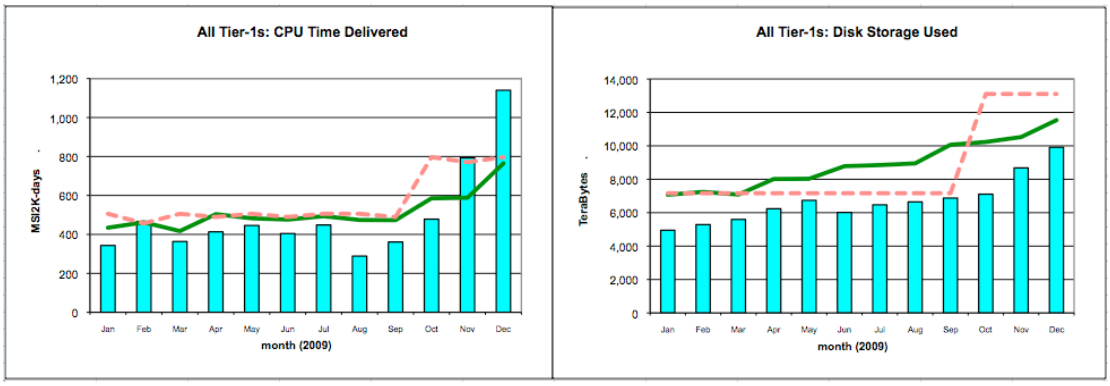


Figure 1d: CPU and disk space used in 2009 at all Tier-1 centres summed together (histograms), compared to resource pledges (dashed pink lines) and installed capacities for ATLAS (solid green lines). For direct reference, the requested CPU resources were 462 MSI2k-days per month till September and 516 MSI2k-days per month from October onwards; the disk request was 7.5 PB till September and 8.6 PB from October onwards.

Tier-1	**Pledge (HS06-days)**	**Available (HS06-days)**	**Used (HS06-days)**	**% Used of available**
CA-TRIUMF	1.29M	1.23M	1.84M	149
DE-KIT	2.57M	2.77M	3.01M	109
ES-PIC	1.19M	1.33M	1.44M	108
FR-CCIN2P3	2.93M	3.29M	2.77M	84
IT-CNAF	1.42M	1.13M	1.78M	157
NDGF	1.45M	1.50M	1.38M	92
NL-NIKHEF/SARA	4.54M	3.27M	2.67M	82
TW-ASGC	2.36M	0.87M	1.30M	149
UK-RAL	2.73M	2.38M	2.04M	86
US-BNL	6.79M	6.85M	5.58M	81
Total	***27.27M***	***24.62M***	***23.79M***	***97***

Table 1: Summary of the CPU pledged, made available and used at each Tier-1 centre in 2009, in HS06-days

Table 1 shows a summary of the CPU pledged, made available and used at each Tier-1 centre in 2009, and their sums. Looking at all Tier sites, ATLAS used in 2009 23.8M HS06-days of CPU at Tier-1 sites and 45.5M HS06-days of CPU at Tier-2 sites, corresponding respectively to 1/3 and 2/3 of the Grid CPU usage. Figure 2 gives the relative distribution of used CPU at Tier-1s and Tier-2s by country.

The ATLAS requirements for the calendar year 2009, including the standard efficiency factor 85%, and considering the 2008 requirements until September and the 2009

requirements (as revised in Summer 2009) from October onwards, were 23.1M HS06-days of CPU at Tier-1 sites and 24.9M HS06-days of CPU at Tier-2 sites.

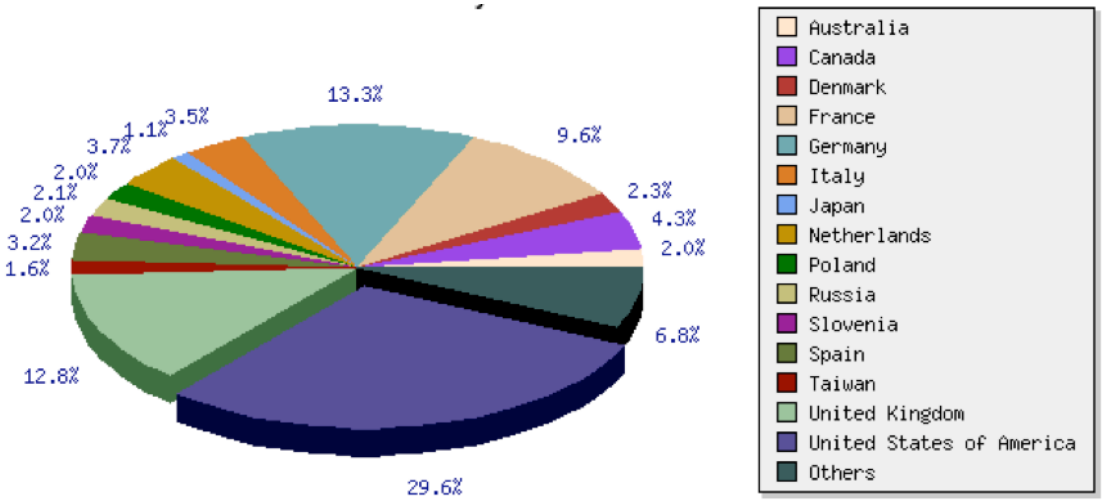


Figure 2: Distribution by country of the CPU used on the Grid in 2009, summed over Tier-1s and Tier-2s. "Others" includes the 15 countries contributing less than 1% each.

ATLAS requirements for disk space were 10.7 PB at Tier-1s and 7.8 PB at Tier-2s until September 2009 (2008 accounting year) and 12.2 PB at Tier-1s and 10.8 PB at Tier-2s afterwards.

The total amount of disk space pledged at Tier-2 sites for ATLAS is 13.3 PB for the 2009 accounting year (i.e. between 1 October 2009 and 31 May 2010). On 3 March 2010, the measured available disk space is 12.2 PB, 4.6 PB of which (38%) are currently occupied. See Section 5 for a breakdown of data types currently stored at Tier-2 sites.

3. Simulation production

The simulation production activities continued throughout 2009. Figure 3 shows the number of instantaneously running jobs, for each "cloud" made of a Tier-1 centre and the associated Tier-2s.

With the MC08 cycle (until Summer 2009) we produced 300M events, mainly at a centre-of-mass energy of 10 TeV, which were reconstructed twice with different software versions (the last time in September 2009).

The MC09 production cycle consisted of the following fully simulated samples:

- 185M events (mostly minimum bias) at a centre-of-mass energy of 900 GeV
- 20M events (minimum bias) at a centre-of-mass energy of 2.36 TeV
- 315M events at a centre-of-mass energy of 10 TeV
- 55M events at a centre-of-mass energy of 7 TeV

In addition, 385M events were produced in the first half of 2009 with the fast simulation program Atlfast-II.

The large statistics of 900 GeV events arises from a combination of factors, such as that initial expectations were for a substantially larger data sample, and samples were produced with different simulation options, for example because of improved

understanding of the material in the detector and for systematic studies for the first ATLAS physics publications. These samples have been very useful for detailed studies of the simulation performance.

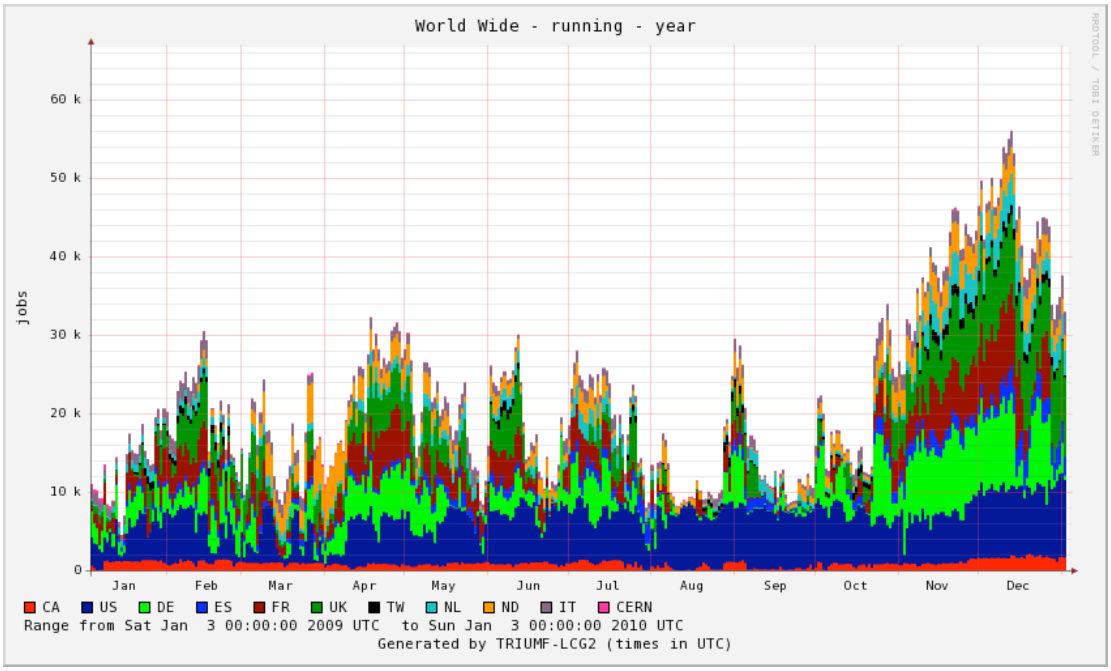


Figure 3: Number of ATLAS jobs running at the same time on the Grid in 2009, subdivided by "cloud".

All MC09 events were re-reconstructed in November-December 2009, and additional events were produced since. In order to maintain consistency between simulated and real data, we plan to re-reconstruct simulated events every time we reprocess real events with improved software, although data from only the two most recent reprocessings are retained.

Figure 4 shows the amount of data produced in 2009 during the MC08 and MC09 cycles. In total we produced in 2009 2.3 PB of simulated events. Some of these events are still useful in 2010, although larger samples of events at a centre-of-mass energy of 7 TeV are now in production with an updated geometry; as these are produced, the MC08 and 10-TeV MC09 samples will be deleted from disk.

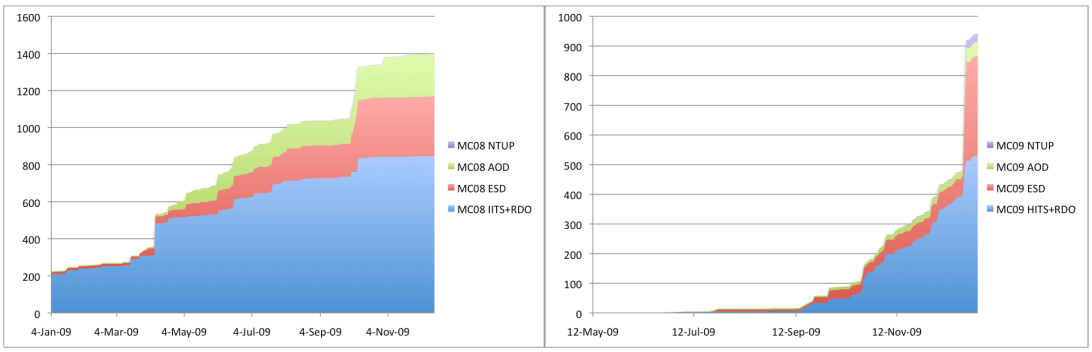


Figure 4: Data produced by the MC08 and MC09 simulation cycles in 2009, before dataset replication. Units are TB. See Appendix A for details of data formats.

4. Collection and reprocessing of real data (cosmics and LHC)

There were two major campaigns of cosmics data-taking, respectively in May-June and October-November 2009. In total, 850 TB of RAW data were collected, and the corresponding amounts of processed data for a total of 1.1 PB, as shown in Figure 5. Cosmic ray data are useful to exercise and test the detector and data acquisition systems. For purposes of calibration and alignment of detector systems, cosmic ray data have periodically been reprocessed with updated versions of the software and calibration constants.

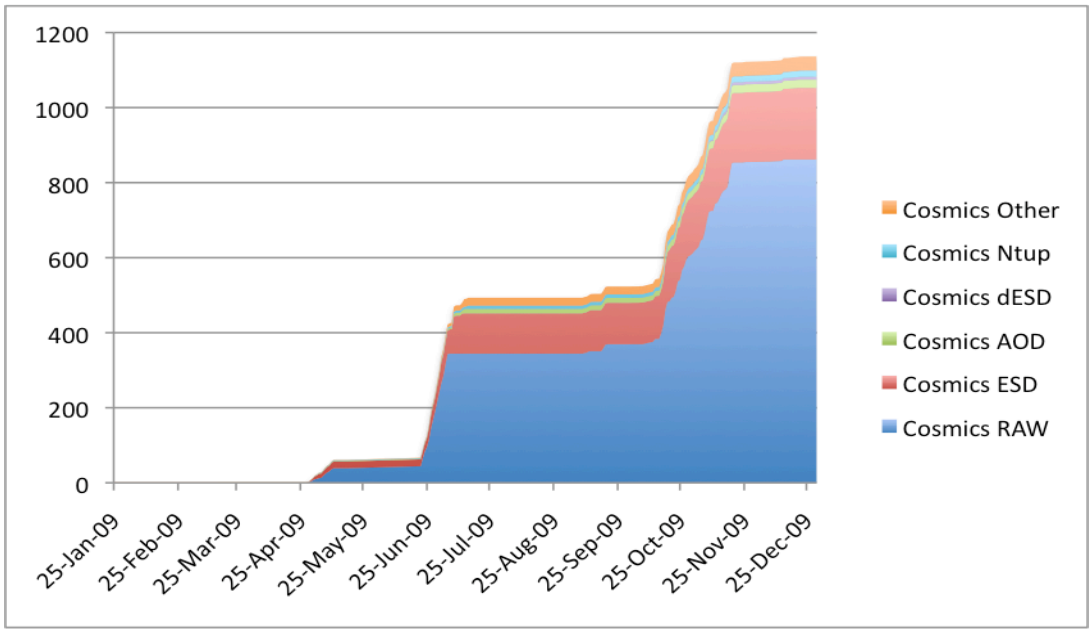


Figure 5: Cosmic ray data taken in 2009, before dataset replication. Units are TB. See Appendix A for details of data formats.

Figure 6 shows the jobs running on the Grid during the reprocessing cycles in 2009. In March and April, 2008 data were reprocessed. Cycles between June and September refer to 2009 data. The December cycle is the first global reprocessing of 2009 LHC data.

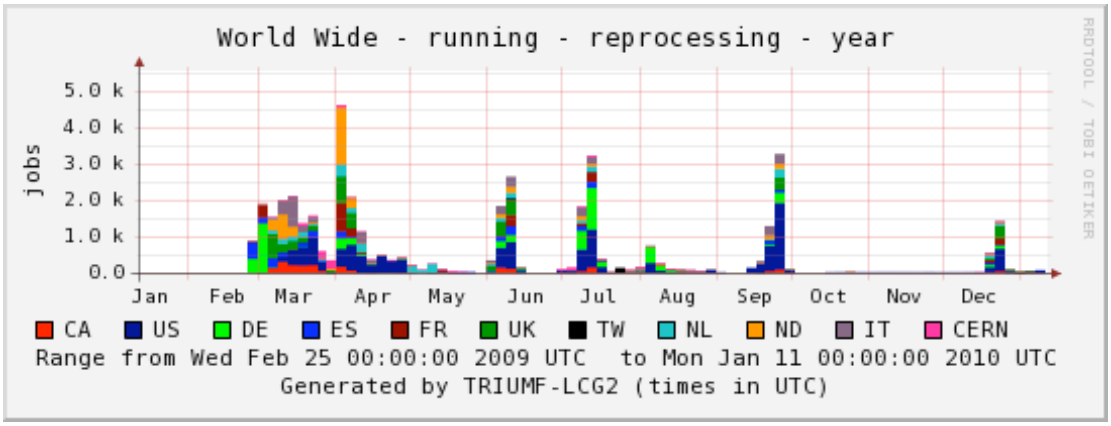


Figure 6: Cosmics and LHC data reprocessing jobs running on the Grid in 2009.

LHC data taking started on 20 November and continued till 16 December. We accumulated 170 TB of RAW data. Including the output of the first global reprocessing at the end of December 2009, the 2009 run produced in total 220 TB of data (see figure 7).

See Appendix A for details of data formats and average sizes for collision events. The large ratio between RAW and ESD sizes in the 2009 beam data-taking comes about because the LAr calorimeter read out several time samplings of cell signals, as well as the fitted pulse characteristics. This resulted in a RAW event size of 5 MB compared to the nominal 1.6 MB per event. The average data rate was kept below 320 MB/s, as the trigger rate was low. As the rate of LHC collisions increases, the trigger rate will increase and the number of samplings will be reduced to keep the average data rate below 320 MB/s. The number of samplings has already been reduced, reflecting the increased confidence in the ROD DSP pulse shape fitting procedure. The increased RAW data size from this source does not impact the event size of derived data (ESD and subsequent stages).

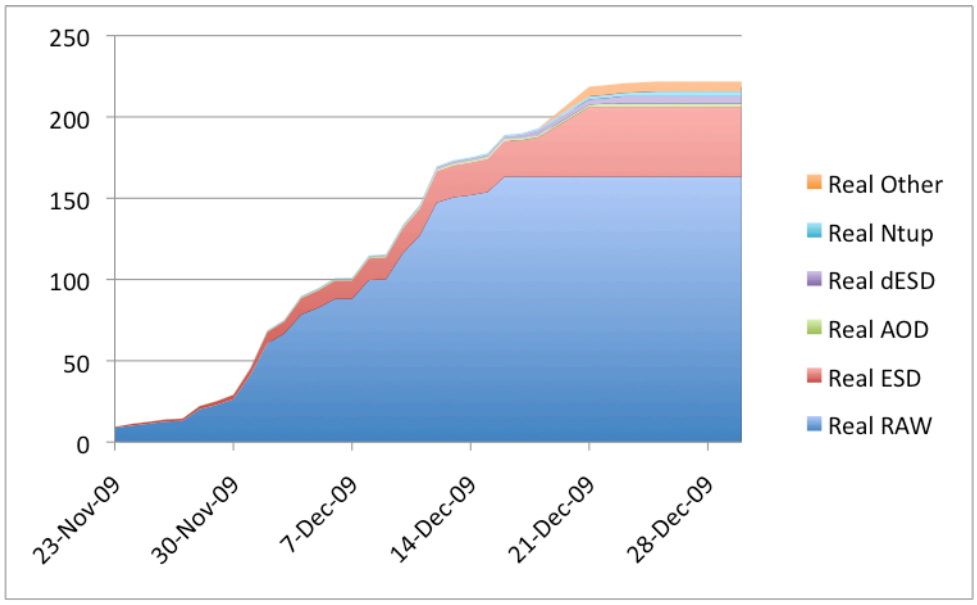


Figure 7: Real data produced by the LHC and reprocessing at the end of 2009, before dataset replication. Units are TB. See Appendix A for details of data formats. "Other" includes data quality histograms that are summed and uploaded to the CERN data quality servers and log files.

Table 2 shows the total amount of data collected or produced in 2009, by data type. Only one copy of the data is considered for table 2, in the same way as for figures 4, 5 and 7. The MC08 simulated data, which turn out to be of little interest for the 2010-2011 run, will soon be removed from disk.

Data (TB)	RAW/HITS	ESD	AOD	dESD	Ntuple	Other
LHC	163	45	2.2	4.9	3.2	6.4
Cosmics	862	191	22	7.7	16.9	37.2
MC08	848	323	234	–	1.5	–
MC09	781	375	104	–	27.9	–

Table 2: ATLAS data produced in 2009, in units of TB. It includes 2 versions of reconstructed data, both for simulated and real data. dESDs were not produced for simulated data.

5. Data distribution and deletion

In 2009 ATLAS had small volumes of real data from LHC, so it was decided to use available disk resources to make access to the data as easy as possible for analysis. Initial data are particularly useful for detector commissioning, therefore the RAW and ESD are the most used data formats. As more events will be collected in 2010, we anticipate that the interest will shift towards more compact data formats, such as the AODs and dESDs.

During 2009 ATLAS developed tools allowing the logging of dataset access times and frequencies, more efficient tools to delete datasets and update catalogues, and tools to run consistency checks between the information in DDM catalogues, SRM interfaces and Storage Elements. In this way it is possible to make the most effective use of available disk capacity until new data become available.

All types of data were distributed in 2009 in more copies than described in the computing model. There were two main reasons for this:

1. Cosmics and simulated data were distributed, prior to the first LHC run, as if they were real data, in order to test thoroughly the data distribution system. Test data were also distributed all the time for the same purpose, but then removed periodically.
2. LHC data are used both for detector commissioning and physics analysis. The RAW data were sent to both disk and tape at Tier-1s, according to Tier-1 shares; in addition, a full copy of RAW data was sent to disk in CCIN2P3, BNL and SARA in order to ease access for detector analysis. ESD data, which is the most used data format at this stage, were sent to all Tier-1s and further distributed to Tier-2s. 18 copies of ESDs, AODs and dESDs were distributed globally to Tier-2s. About 1 PB of data (including all replicas) were generated from the 2009 LHC run.

The computing model foresees that two copies of the latest version of ESDs and AODs will be distributed and stored on disk across the Tier-1 sites, and ten copies of AODs and dESDs will be distributed across the Tier-2s. The copies distributed according to the computing model are marked in the data catalogues as “primary”; the additional copies are marked as “secondary”. Secondary copies will be progressively and automatically

removed from disk as soon as disk space at that site and for that data type will become scarce, starting from the least used data.

ATLAS disk storage space is organised using the concept of "space tokens"; a space token is the storage area that is allocated to a given activity, with a given quota. In this way it is possible to reserve storage for central distribution of data, such as the output of simulation production, Tier-0 processing or reprocessing of real data, separately from the activities of analysis groups and users.

Table 3 shows, for all Tier-1s added together, the space occupied by different types of data and the free space in the associated space tokens:

- DATADISK and MCDISK hold respectively real (cosmics and LHC) data and simulated data; they currently include the additional (secondary) copies described above.
- GROUPDISK is the area allocated to ATLAS activity groups to store and share their derived data. GROUPDISK usage is still low, at present, in absence of large collision data samples, but will grow sharply as soon as the 2010 LHC run starts.
- Buffers include the space reserved for production activities and for the output of analysis jobs. As data there is by nature transient, they also act as a "reserve" in case there is urgent need to add capacity to one of the other space tokens. In addition, there are buffers in front of tape systems; they will normally have a low occupancy, except during data-taking periods when the throughput rate is large.

Storage area (Tier-1s)	**Used (TB)**	**Free (TB)**	**Installed (TB)**	**% used**
DATADISK	2626	2842	5468	48
MCDISK	4494	1251	5745	78
GROUPDISK	119	483	602	20
Buffers	415	866	1281	32
Total	*7654*	*5442*	*13096*	*58*

Table 3: Data stored on 3 March 2010 in each space token and available disk space for all Tier-1s together, after initial cleaning of disk space in preparation for new data. Installed capacities in this table are measured through SRM queries.

Table 4 shows the same information, but for Tier-2 sites. The bottom rows of tables 3 and 4 can be misleading, unless one considers that buffers are needed for production operations, and part of the disk capacity that became available at the end of 2009 and early in 2010 cannot be filled in immediately as we are about to start a long data-taking period in 2010. As explained above, we have already started clearing away all secondary

copies of old data to make room for new 2010 data to come. Note that these effects give a different picture from that of figure 1 because the preparation for 2010 data is now well underway; this snapshot is taken 2 months later than the last entries of figure 1.

Storage area (Tier-2s)	**Used (TB)**	**Free (TB)**	**Installed (TB)**	**% used**
DATADISK	1239	2330	3569	34
MCDISK	2687	1259	3946	68
GROUPDISK	278	1790	2068	13
Buffers	401	2209	2610	15
Total	*4605*	*7588*	*12192*	*38*

Table 4: Data stored on 3 March 2010 in each space token and available disk space for all Tier-2s together, after initial cleaning of disk space in preparation for new data. Installed capacities in this table are measured through SRM queries.

6. Analysis jobs and data access

During 2009 we distributed additional copies of real and simulated data on the Grid with respect to the computing model, in order to ease user access to the early data. This was possible as the 2009 run started only in November, rather than in September as anticipated at the time of the resource requests. In this way it was possible to access ESD data at Tier-2 sites. ESDs were indeed the most popular data format for the analysers of 2009 LHC events, as they contain the most complete information about the reconstructed event quantities.

Dataset type	Number of accesses	Number of sites
RAW	11480	13
ESD	199485	66
AOD	6438	64
dESD and Ntuple	11274	51

Table 5: Number of accesses to datasets of LHC data of different types on the Grid and number of sites involved, between 20 November 2009 and 12 February 2010.

Table 5 shows the number of accesses to datasets of different types on the Grid and the number of sites involved; it is evident that the wide distribution of ESDs resulted in a wide use of Grid tools to submit jobs to all sites holding the data. Figure 8 shows the

distribution of analysis data access requests on the Grid. In total 68 sites received jobs or data access requests; whereas it is not surprising that BNL had the largest share, as it is a large Tier-1 site with a complete set of data on disk that allows user access, it is also comforting to see a healthy distribution of user tasks also to many other sites; in fact, the third, fourth and sixth sites are Tier-2s, not Tier-1 centres.

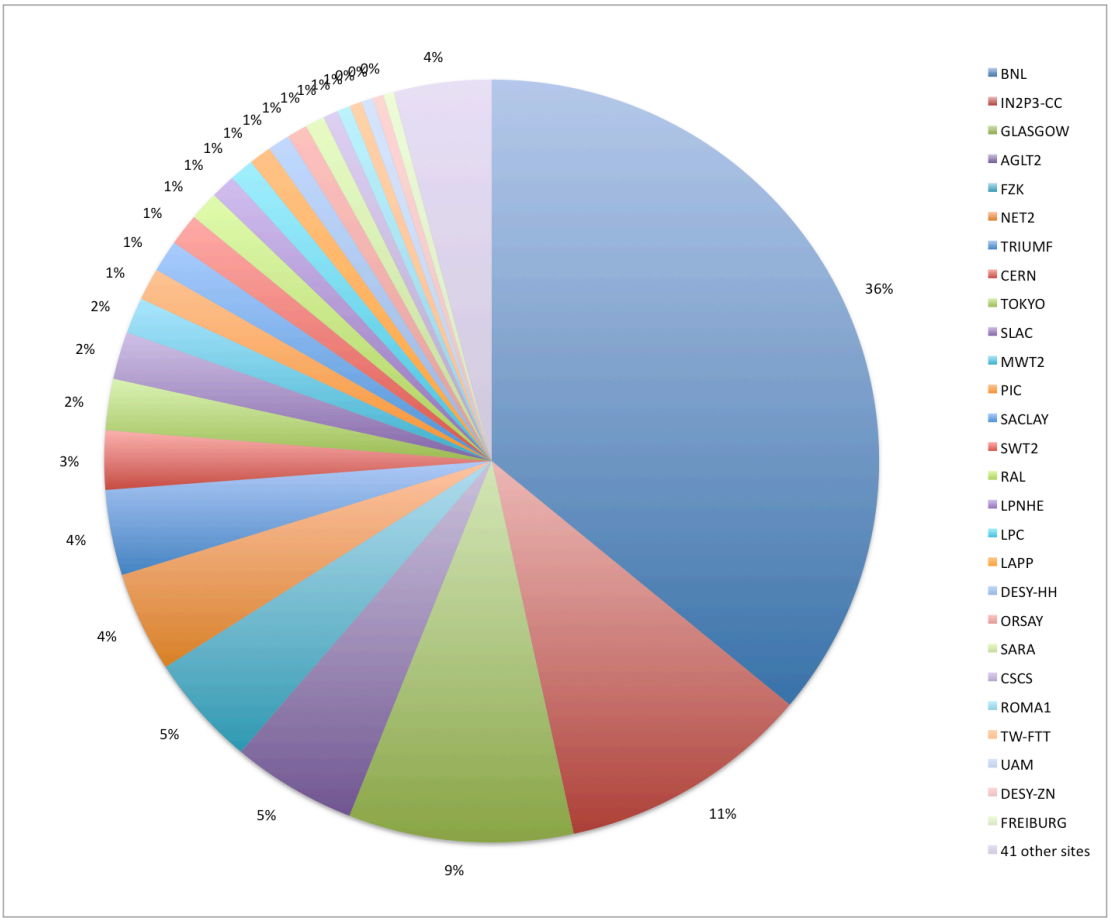


Figure 8: Distribution by site of dataset access requests for LHC data of all formats, in the period between 20 November 2009 and 12 February 2010.

The analysis tasks were run by several hundred ATLAS physicists. A large fraction of the datasets were analysed by a large number of physicists; table 6 shows the number of distinct people that accessed the datasets containing 2009 LHC data, cosmics and simulated data between November 2009 and February 2010. As expected, the interest in the cosmics runs decreased sharply as soon as LHC data became available.

Data type	RAW	ESD	AOD	dESD
LHC data 2009	102	455	527	725
Cosmics 2009	33	87	30	21
MC09	–	523	590	–

Table 6: Number of distinct users that accessed ATLAS data on the Grid (by data type) in the period between 20 November 2009 and 23 February 2010.
Simulated HITS (the equivalent of RAW for real data) are kept on tape and not made generally available for analysis. No dESDs were produced for simulated events.

7. Conclusions

ATLAS has made full use of the computing resources made available at all Grid centres in 2009, including opportunistic usage of idle CPU cycles. We have distributed additional copies of RAW and ESD data produced during the short LHC run at the end of the year in order to ease access to the complete information for first LHC events to all members of the Collaboration. Tools have been developed to remove from disk these additional copies when disk space is needed for new and more popular data. Analysis on the Grid is in full swing, with several hundreds of people routinely submitting jobs through the pAthena and GanGA interfaces.

Acknowledgements

We would like to thank the Funding Agencies and the management of CERN and of all Tier-1 and Tier-2 computing centres for providing the computing resources that are necessary for the success of the ATLAS experiment.

Appendix: ATLAS data formats

The physics event store holds a number of successively derived event representations, beginning with raw or simulated data and progressing through reconstruction into more streamlined event representations suitable for analysis. Constituent components are described in the following paragraphs.

The data emerging from the event filter to enter the offline system are streamed according to trigger signatures. This ensures that the stream definition is immutable under recalibration and reprocessing. The streaming is inclusive, i.e. a given event may occur in more than one stream. The estimated stream overlap is below 10%.

RAW Data: RAW data are the output of the Event Filter (EF, the final stage of the High-Level Trigger), and input to reconstruction. The average event size is 1.6 MB, with an output rate of 200 Hz. Events arrive from the Event Filter in "bytestream" format, reflecting the format in which data are delivered from the detector, rather than in any object-oriented representation. Events are transferred from the EF to the Tier-0 in files. Each file contains events belonging to a single run, trigger stream and luminosity block (typically 2 minutes long), but the events in each file are not consecutive nor ordered.

Event Summary Data (ESD): ESD refers to event data written as the full output of the reconstruction process. The average ESD event size is 0.8 MB for real data and 1.1 MB for simulated events. Its content is intended to make access to RAW data unnecessary for most physics applications other than calibration or re-reconstruction. ESD has an object-oriented representation, and is stored in POOL ROOT files.

Derived ESD Data (dESD): dESD are skimmed events starting from ESD datasets, useful for detector and combined performance studies. There are several streams of dESD, with different selection criteria depending on the detector or performance group that requested them. Whereas the average single event size is close to ESD events, because of the selections the total data volume of dESD datasets will be close to the total volume of AOD datasets.

Analysis Object Data (AOD): AOD is a reduced event representation, derived from ESD, suitable for analysis. It contains physics objects and other elements of analysis interest. The average AOD event size is 150 kB for real data and 180 kB for simulated events. It has an object-oriented representation, and is stored in POOL ROOT files.

Tag Data (TAG): TAG data are event-level metadata — thumbnail information about events to support efficient identification and selection of events of interest to a given analysis. Their average event size is 1 kB. To facilitate queries for event selection, TAG data are stored in a relational database, as well as in ROOT files.

Simulated Event Data (HITS and RDO): HITS datasets contain the output of the full Geant4 simulation of the ATLAS detector. They are stored in POOL ROOT files. Digitised events may be stored in POOL ROOT files (RDOs) or in bytestream format for trigger studies or for emulation of data coming from the Event Filter. Simulated events are often somewhat larger than RAW events (approximately 2 MB in size), in part because they usually retain Monte Carlo "truth" information.

Ntuples and Histograms: at every processing step, ntuples and histograms are produced to monitor the processing and data quality. They are usually transferred to CERN, and histograms are summed and loaded to the data quality servers. There are also calibration ntuples, which contain very reduced subsets of selected events and are used for calibration and alignment studies.

References

[1] http://lcg.web.cern.ch/LCG/accounting/Tier1/2009/december-09/Master_accounting_summaries_December2009.pdf

[2] http://lcg.web.cern.ch/LCG/Resources/WLCGResources-2008-2013-20APR2009.pdf

[3] https://twiki.cern.ch/twiki/bin/view/LCG/SamMbReports

Appendix C: CMS Resource Utilization in 2009

Introduction

The approved planning developed in the spring of 2009 called for an initial 6 month running period starting in October and continuing until April of 2010. This was to be followed by a second period from April 2010 until the end of the year. During the first period CMS had expected to collect 10-20 pb-1 of high energy collision data. The recovery from the 2008 incident was longer and more gradual than expected and in the first period CMS collected 10 inverse micro-barn of collision data at a combination of injection energy and 2.3TeV. In addition to the small sample of collision data, CMS completed a second Cosmic Commissioning run (CRAFT09) and a substantial cosmic sample was collected during the beam running period. CMS expects that during the first 6 months of the 2010 run we will execute a program that is similar to the first period of 2009, in terms of data volumes, processing passes, and general start-up challenges.

During the 2009 run, CMS collected roughly 7M events in the minimum bias data sample that were declared good for physics. This is approximately 1% of the total physics data anticipated in the first running period. This is reflected in the utilization of the Tier-0 farm, which is not saturated by cosmic running, and in the utilization of the Tier-1 centers for data reprocessing. The number of cosmic events collected during the dedicated run and in combination with the beam running exceeds 600M events, but the raw event size and processing time required are both small. It is challenging to draw good conclusions from the resource utilization in 2009, because it is driven primarily by commissioning activities and not from steady state running at the anticipated rates.

In the resource utilization descriptions that follow we rely on a variety of sources of information. The CERN Tier-0 and CAF are accessed with local batch queue submission, so CERN Lemon monitoring is used to document the processing time consumed during 2009. For grid accessible facilities at the Tier-1s and Tier-2s we use the WLCG accounting portal as the primary source of information. The accounting portal is cross checked with the information from the CMS dashboard. At the time of this document, the last available accounting report for Tier-2s is December 2009, so for consistency we plot all utilization for 12 months of 2009. Processing utilization is shown monthly. Storage utilization is shown as a sum total for 2009.

Tier-0 and CAF

The CPU utilization as a percentage and the number of jobs submitted to the queue are shown in Figures 1 and 2. The cluster labels were changed in the summer so information on utilization begins in August. The late summer and early fall were dominated by Tier-0 commissioning tests. In 2009 CMS produced a “soup” of realistic high energy events and fed them from the Tier-0 through the Repacking steps into a realistic number of primary datasets and ran the events through the prompt reconstruction system. A custom Pythia configuration was employed to generate a mixture of interesting events that could reliably pass the higher level triggers in one job.

The peaks in October are related to the Cosmic tests during 2009 and the collision data is visible during November. The total job slot resources are occupied for individual days as can be seen by the blue periods in figure 2, which indicate pending jobs. The total resources are not completely occupied when we average over a week, which is a demonstration that the event rate and complexity achieved in 2009 was lower than expected.

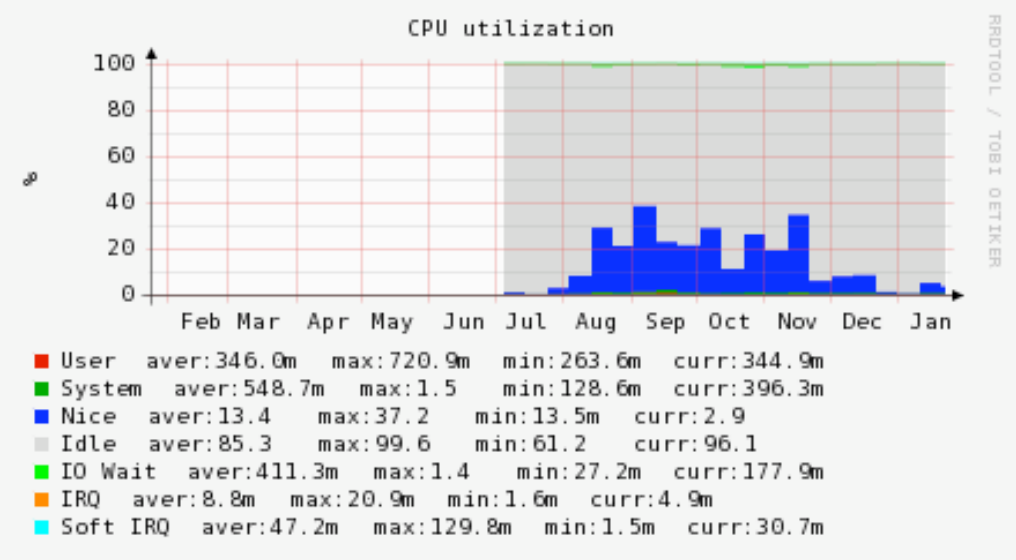


Figure 1: The percentage of the Tier-0 computing capacity used is plotted averaged over a week since July. The blue "Nice" allocation is resources used by normal batch jobs from the CMS production user.

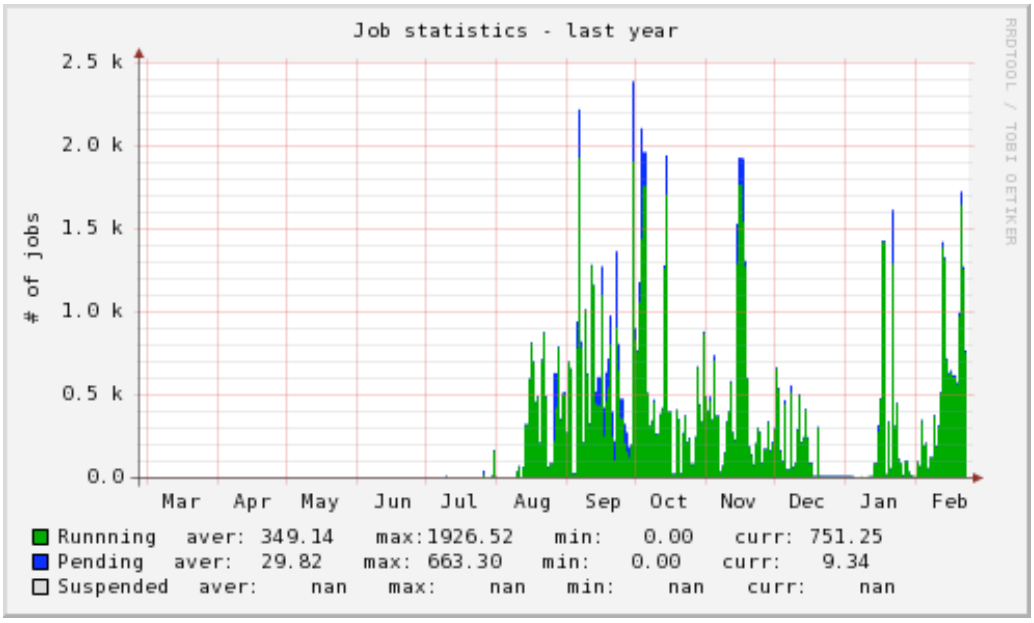


Figure 2: The number of jobs submitted to the CMS Tier-0 batch farm is plotted by day since July. Green indicates running jobs and blue indicates the jobs are waiting for processing resources.

The CAF is primarily intended for low latency commissioning, calibration and analysis work. Detector calibration with data has been restricted to calibrations that can be done with cosmics and calibrations that benefit from the limited minimum bias samples. Many of the calibration workflows are waiting for significant data samples. The lack of large data volumes and the desire of the collaboration to commission the distributed infrastructure for analysis is reflected in the low average utilization of the CAF batch slots. The CAF has less than 700 batch slots, but the average utilization in 2009 was 13% on average. As is visible in Figure 3 the CAF resources have been fully utilized in particular periods. The total number of running jobs is limited by the available batch slots, but also in the number of running jobs each person can submit. In order to ensure low latency access for others each individual user is limited to 100 running jobs. The CPU efficiency for CAF jobs averaged roughly 52%. This is lower than would be expected from regular analysis jobs, but factors in data intensive calibration workflows. The expectation is that with a larger data sample and an active program of work in calibration, CMS will be resource constrained.

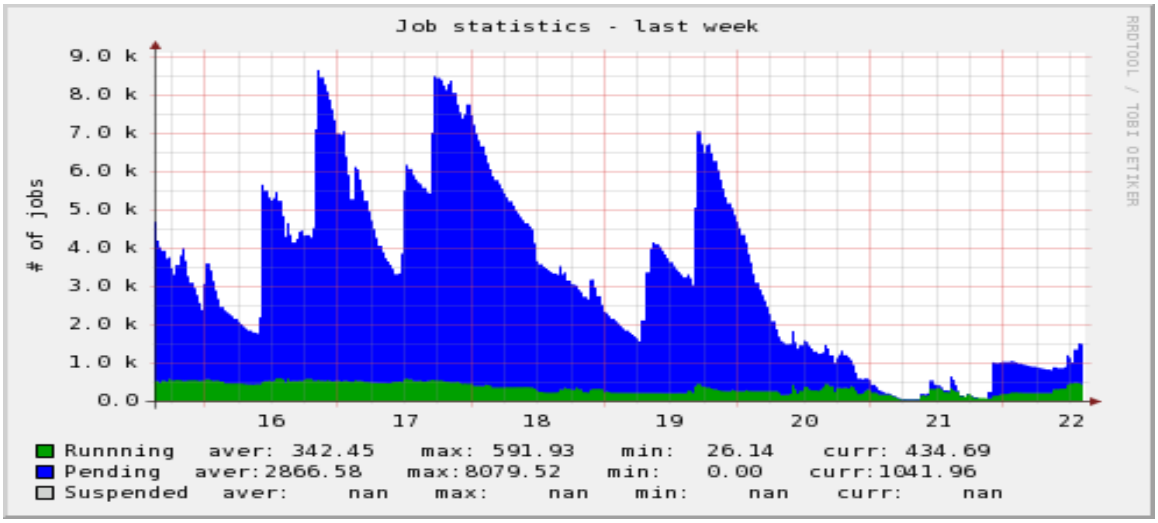


Figure 3: Jobs submitted to the CAF by commissioning and analysis users during a week in April, 2009, are plotted. Green is consumed job slots and Blue indicates waiting jobs.

The CAF disk is currently 1.2PB split between the calibration program, detector commissioning, and physics: 20%,40%, and 40% respectively. The regular cosmic running combined with test workflows have made effective use of the storage with it completely filling once during server maintenance.

Tier-1 Resources

In the CMS model the Tier-1s primarily archive and serve data and simulation, while processing and skimming the stored samples. The lower rate data, both in terms of the percentage of interesting physics data and the total volume, results in a lower utilization of the Tier-1s for data reprocessing and collision data storage, but the Tier-1s were heavily used for storing simulation throughout 2009 and the computing operations teams exercised the Tier-1 facilities with repeated commissioning reconstruction passes to reach a good level of readiness.

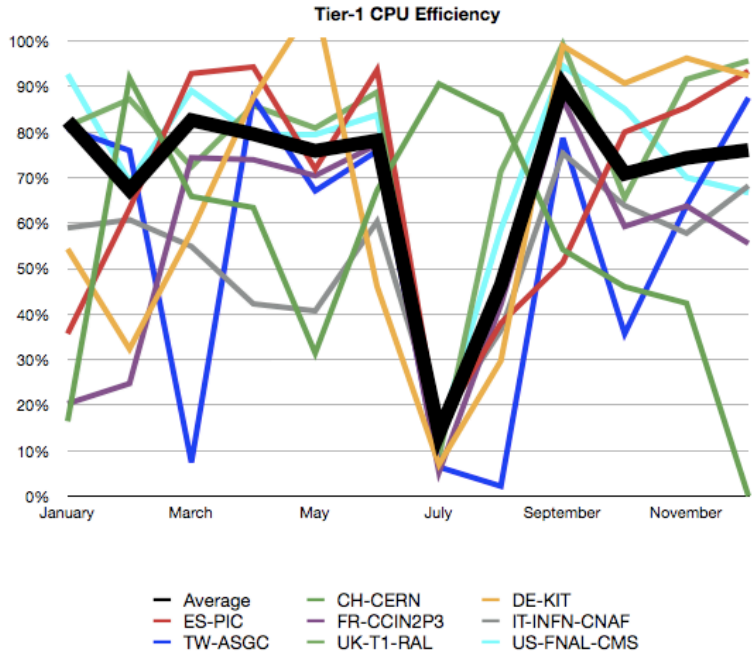


Figure 4: CPU Efficiency for jobs on CMS Tier-1 resources are plotted by month. The large black line is the weighted average.

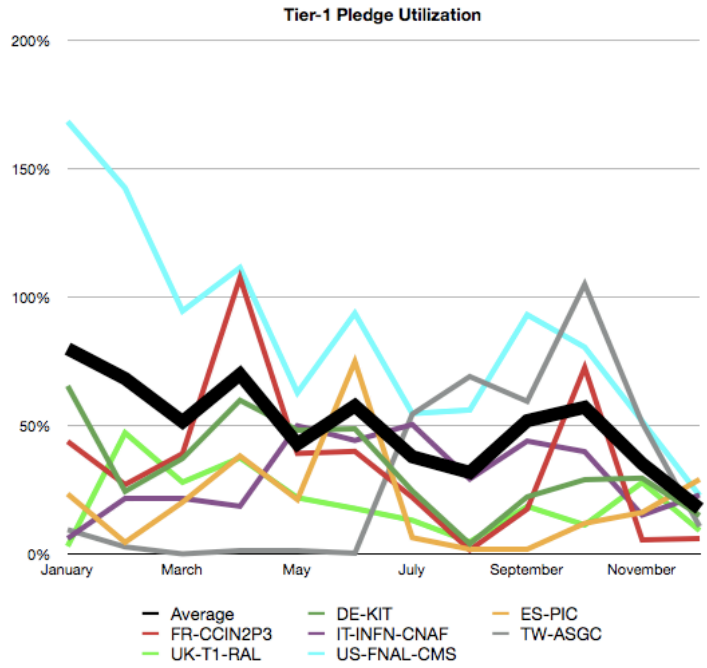


Figure 5: The percentage of the processing pledge utilized by CMS at the Tier-1 computing centers is plotted monthly. The black line is the weighted average across Tier-1 sites.

Figure 4 shows the CPU efficiency as a function of month at the Tier-1 centers. As is clearly visible from the plot there is a big reduction in the summer, which was related to an IO problem identified in the software, but the average of the remaining points is generally within the window expected for organized processing. The average value

shown in black is the total number of CPU hours delivered to CMS divided by the total number of wall clock hours delivered to CMS. This will weight the efficiency of the largest sites higher in the average.

Figure 5 has the average pledge utilization as a function of month in 2009. The pledge utilization is calculated by using the number of normalized wall clock hours extracted from the accounting portal and comparing to the number of processing resources reported as being available to CMS as of October 2009 multiplied by the number of hours in a month. As can be seen in the plot, the highest level of activity is at the beginning of the year when there was a large scale reprocessing of the cosmic runs taken in 2008. There are local fluctuations throughout 2009 as sites participate in commissioning activities and data operations executed a limited number of low latency simulated event productions using the Tier-1s. The total volume of data in 2009 could be re-processed quickly so the period of lowest activity is as the commissioning activities ramp down at the end of the year and the reprocessing of new events had not ramped up to compensate. The total weighted average by delivered CPU capacity is shown in black and is roughly 50% over the course of the year.

Figure 6 shows the disk and tape utilization for CMS and the other LHC experiments at the Tier-1s plus CERN during 2009 directly from the WLCG accounting reports. By the end of 2009 roughly 7PB of disk and 18PB of tape were used. The volume of data on disk uses nearly all the 2009 pledge of 6.5PB at Tier-1s plus CERN capacity and the tape is close to the total 2009 tape request of 11.9PB at Tier-1 plus 7.3PB at CERN for 19.2PB total. The Tier-1 storage is used for archiving data and simulation. During 2009 CMS produced 1.7B simulated events primarily using Tier-2 processing resources, but the events were archived and served from Tier-1 centers. In the CMS computing model only 10% of the simulated events are expected to be staged on disk at any given time, the fraction in 2009 is higher.

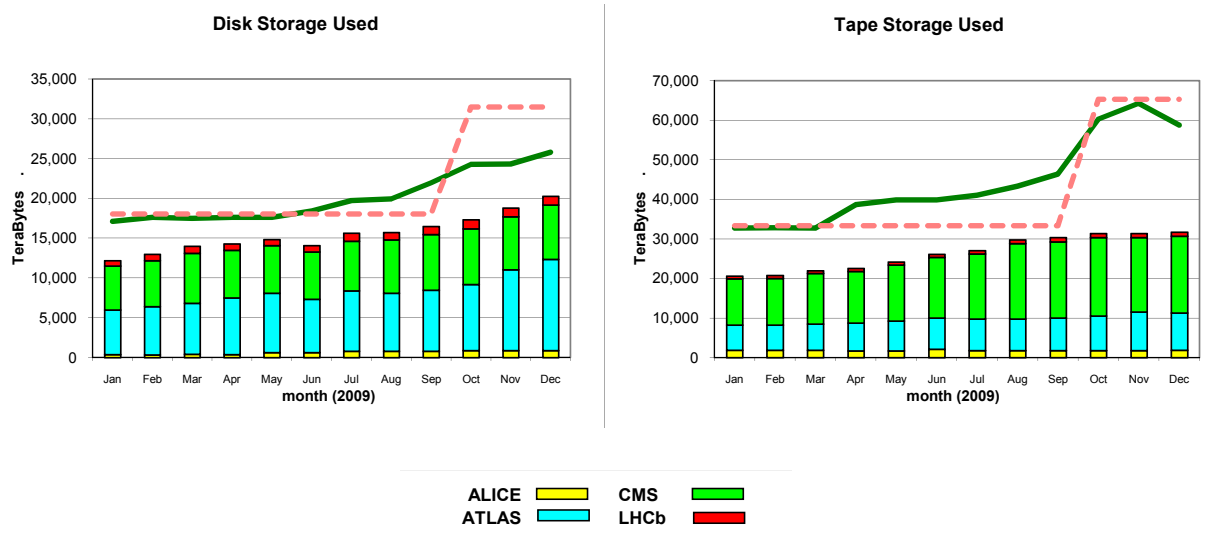


Figure 6: Disk and Tape utilization during 2009 for the 4 LHC experiments at CERN and the Tier-1 centers is taken from the WLCG Tier-1 Accounting Report.

A program of cleaning extra copies from the archives and remove deprecated simulated events is currently ongoing.

Tier-2s

The utilization at the Tier-2s is driven by simulated event production and user analysis that are more loosely coupled to the rate of collision data than the Tier-1 reprocessing. CMS observed that in 2009 the rate of MC production requested and used for commissioning and analysis preparation was similar to what had been anticipated if 2009 had a long data run. As expected, the rate of analysis submission in 2009 peaked shortly after the data run, but the level of submissions is higher than would have been predicted by the volume of data and is driven by physicist enthusiasm. The CPU efficiency and utilization come from the WLCG accounting portal. The Tier-2s are accounted by site, country, or federation depending on the location. Each line of the plot corresponds to an accounting group.

Figure 7 and 8 show the CPU utilization and pledge utilization at the Tier-2 centers. The CPU utilization has a wide spread, but the weighted average is roughly what was predicted from a mix of organized processing jobs and user analysis jobs. One site is reporting equal wall clock and CPU times, which is an indication of a reporting issue and they appear as a horizontal line at 100%.

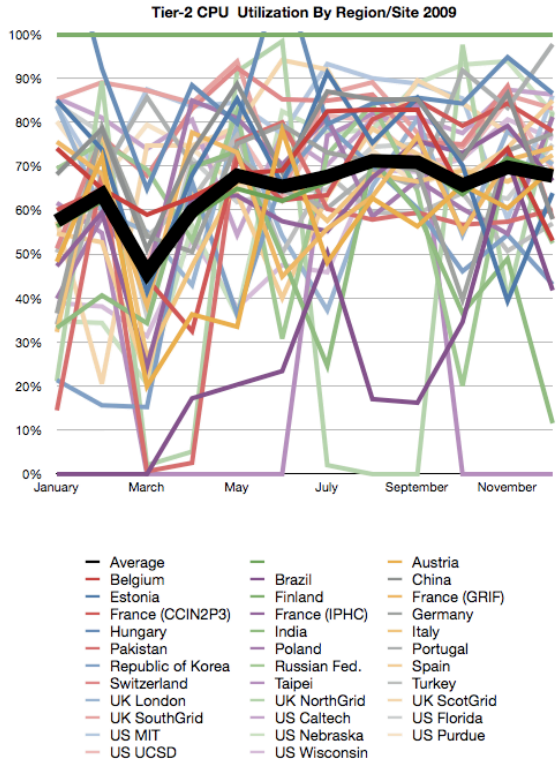


Figure 7: Tier-2 CPU Efficiency is plotted for all accounting groups monthly for 2009. The black line is the average across the Tier-2s

The weighted average of pledge utilization has a wide range, dipping to 50% in the first half of 2009 and growing to more than the 2009 pledge by the latter part of the year. There are a couple of items that contribute to the values. In the WLCG accounting portal the number of processing hours expected of the Tier-2s is the number of hours in a month multiplied by an efficiency factor of 0.6. When calculating the pledge utilization the number of CPU hours delivered is divided by the possible resources including the efficiency factor. As can be seen in the CPU efficiency, we are currently averaging above 0.6, so the efficiency factor may be too conservative. Additionally the possible resources were estimated using the resources listed as available to CMS in October; as Tier-2s deploy additional resources CMS will be seen as being delivered more than the pledge. Figure 9 below shows the number of running jobs at the Tier-2s over the course of the year from the dashboard. The increase in utilization can been seen in the running job plot as well.

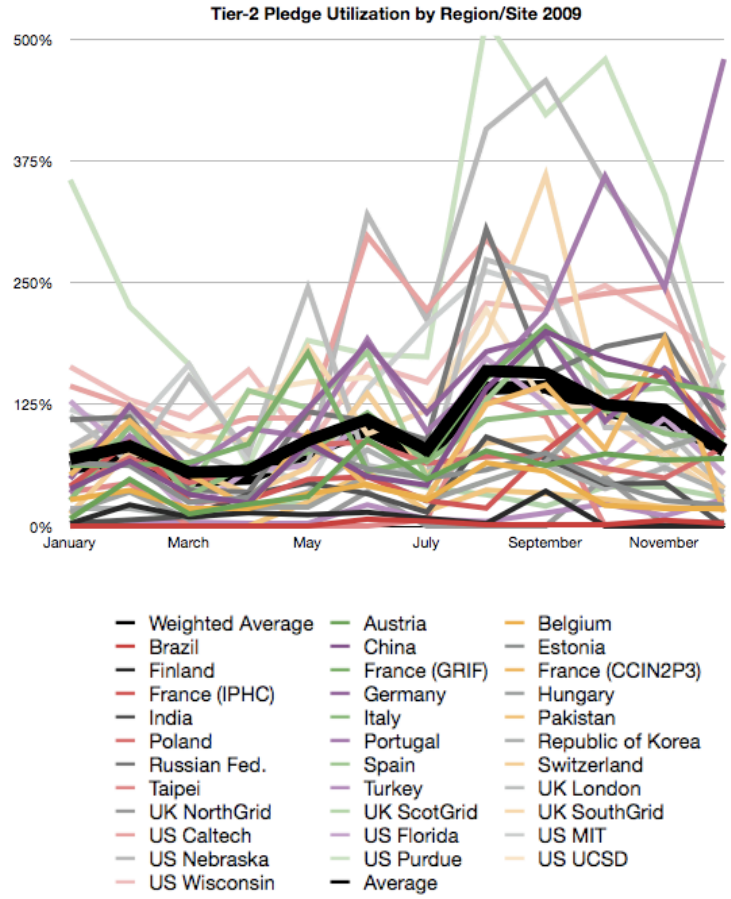


Figure 8: The pledge utilization across Tier-2 sites in 2009 is plotted monthly for CMS by Tier-2 accounting group. The black line is the average weighted by the size of the Tier-2s.

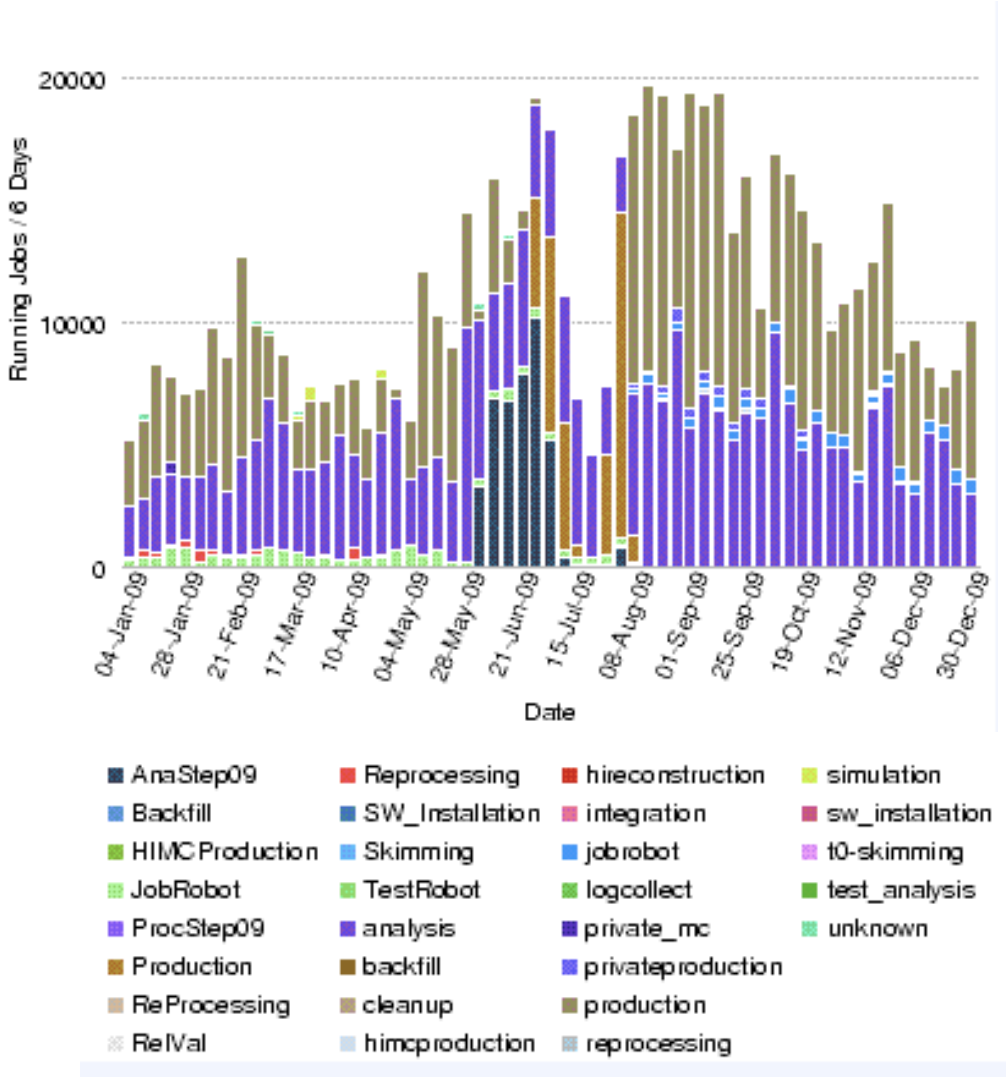


Figure 9: The number of running CMS jobs at Tier-2 sites is plotted over 6 day intervals. Brown is production jobs and purple is analysis jobs from users. Black is jobs from the June scaling activities.

Table 1 shows the disk space used by CMS at the Tier-2 centers for all types of data tracked in the data management system. These numbers will not include individual user files, which are expected to occupy 40-80TB per nominal Tier-2 site. The total value used is very close to the 2009 disk request of 3.7PB, but lower than the total amount of disk available to CMS at the end of 2009. The total volume disk currently available is very close to the original 2010 target and is expected to be fully utilized.

	Used (TB)	Available (TB)
T2_AT_Vienna	92.6	180
T2_BE_IIHE	110.1	240
T2_BE_UCL	110.2	150
T2_BR_SPRACE	4.9	9
T2_BR_UERJ	50.9	200
T2_CH_CSCS	112.4	200
T2_CN_Beijing	40.6	100

	Used (TB)	Available (TB)
T2_DE_DESY	135.5	230
T2_DE_RWTH	118.0	320
T2_EE_Estonia	46.5	80
T2_ES_CIEMAT	92.7	250
T2_ES_IFCA	93.2	190
T2_FI_HIP	60.13	210
T2_FR_IN2P3	164.1	230
T2_FR_GRIF	132.5	180
T2_FR_IPHC	108.9	210
T2_HU_Budapest	50.4	90
T2_IN_TIFR	35.4	450
T2_IT_Bari	107.6	130
T2_IT_Legnaro	117.1	300
T2_IT_Pisa	127.3	190
T2_IT_Rome	61.0	160
T2_KR_KNU	65.4	191
T2_PK_NCP	0.8	70
T2_PL_Warsaw	1.8	40
T2_PT_Lisbon	22.5	200
T2_RU_IHEP	1.4	8
T2_RU_ITEP	1.5	80
T2_RU_INR	1.3	46
T2_RU_JINR	84.8	187
T2_RU_PNPI	0.9	10
T2_RU_RRC	12.0	195
T2_RU_SINP	6.9	123
T2_TR_METU	63.2	190
T2_TW_Taiwan	99.3	110
T2_UA_KIPT	10.5	125
T2_UK_London	74	120
T2_UK_SouthGrid	96	210
T2_US_Caltech	174.4	400
T2_US_Florida	133.0	400
T2_US_MIT	98.6	400
T2_US_Nebraska	194.3	400
T2_US_Purdue	236.5	400
T2_US_UCSD	151.6	400
T2_US_Wisconsin	67.1	400
SUM	3569.83	9004

Table 1: Tier-2 Disk Utilization by CMS and Available Disk Resources in TB.

Transfers

The CMS transfer system was heavily exercised during 2009. The transfers had contributions from normal production and analysis traffic with a continued strong contribution from debugging traffic and commissioning tests. The rate from CERN to the Tier-1 facilities during 2009 is shown in Figure 10, averaged over a week. The week average is lower than anticipated in the planning due to the total volume of data, but the 2009 data taking in November and December is visible in the plot. The volume of data sent to each Tier-1 site during the 2009 collision run matches the the relative fraction of the pledged contribution, with the exceptions of ASGC and IN2P3. ASGC had issues with availability of storage media for the 2009 data period and IN2P3 received a higher relative share.

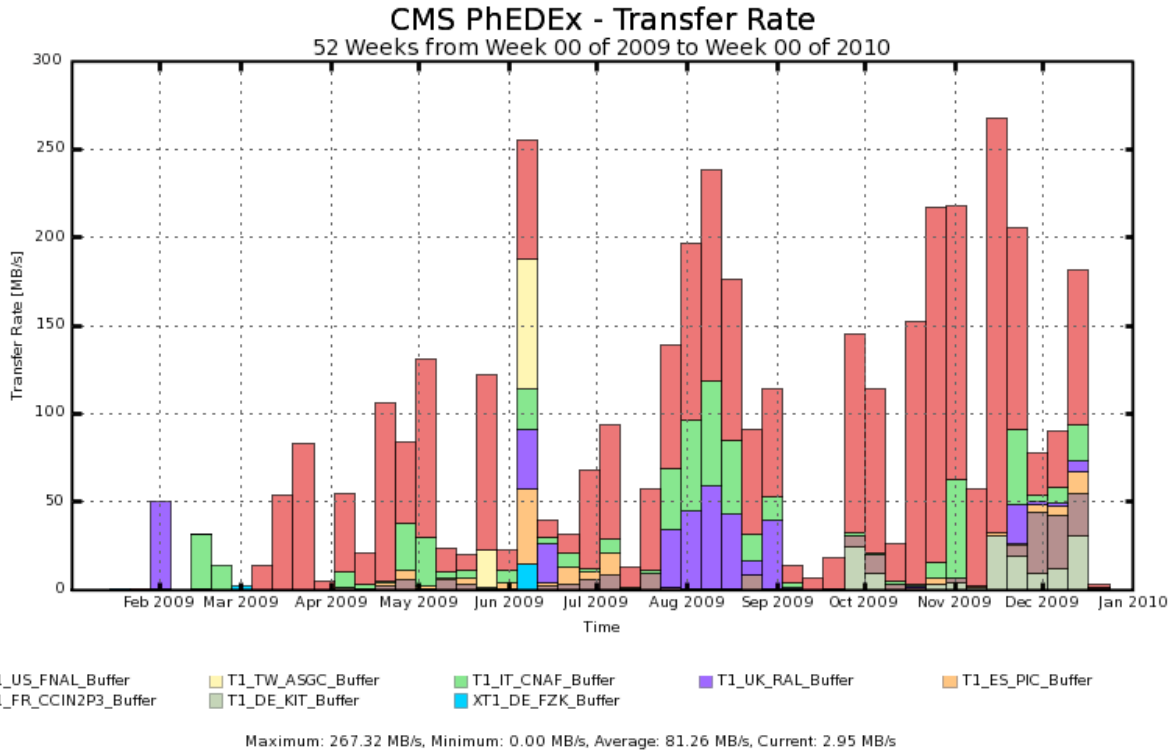


Figure 10: Traffic from CERN to the Tier-1 sites during 2009 is plotted averaged over weeks during 2009 in MB/s.

The total CMS traffic between sites is shown in Figure 11. 90 sites received data through PhEDEx in 2009 and the average weekly transfer rate was nearly 1200MB/s. The peak week averaged close to 2GB/s. The total volume of data transferred in 2009 by PhEDEx adds to more than 35PB, which is shown in Figure 12. As expected the Tier-1s have the highest percentage of received data, but there is a good balance.

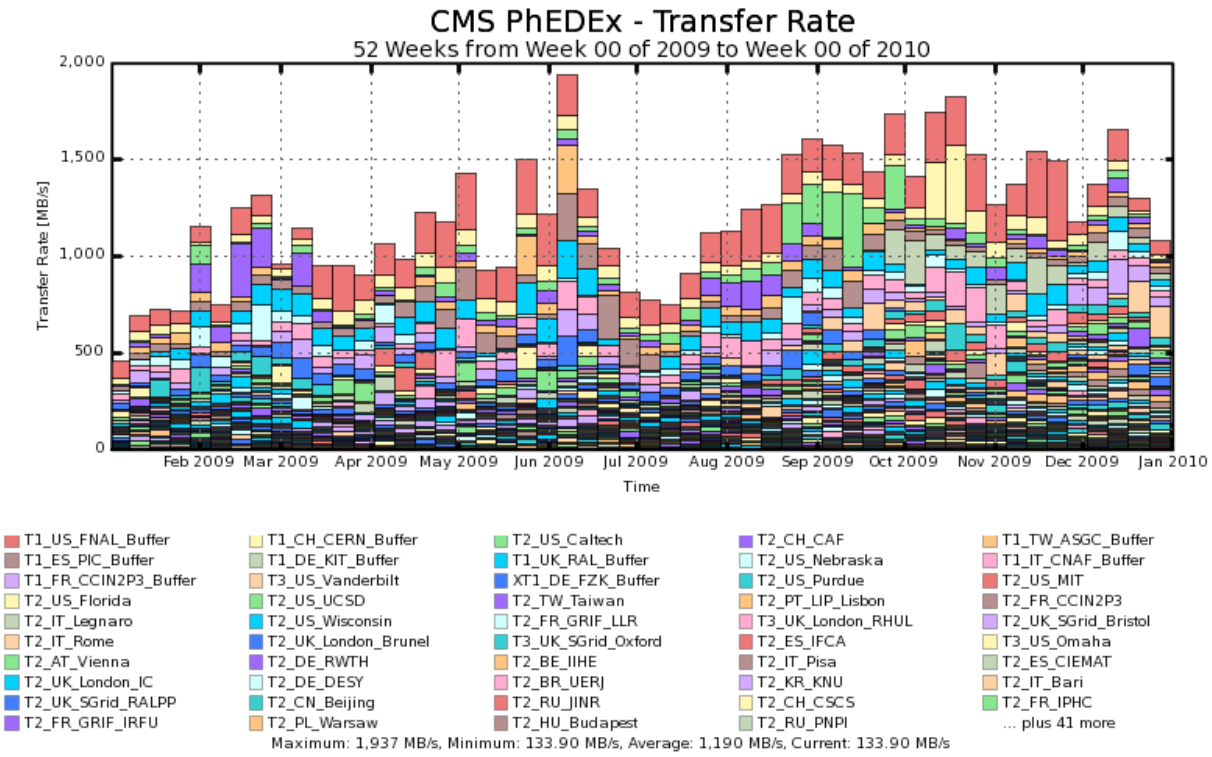


Figure 11: The total rate of data transferred by PhEDEx between all CMS sites averaged over weeks in 2009 is plotted in MB/s.

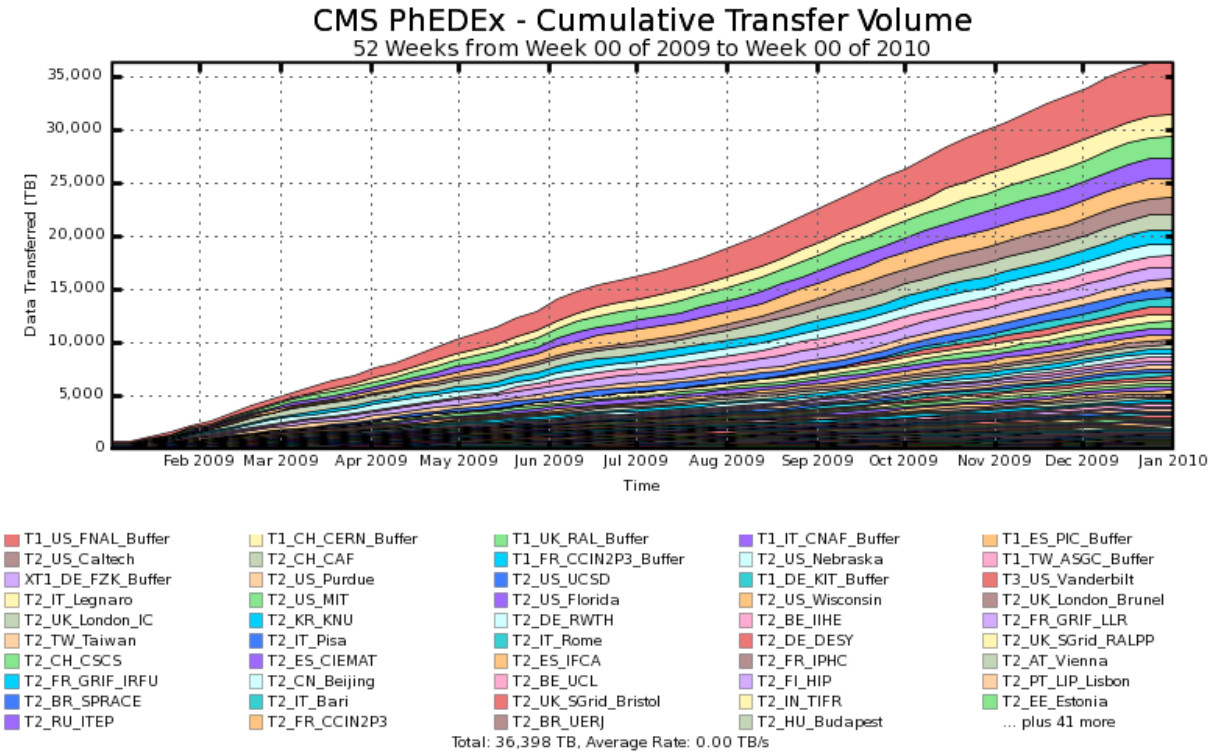


Figure 12: The aggregate volume of data moved by PhEDEx in 2009 between all CMS sites is shown in TB.

LHCb Grid Computing Resources usage in 2009

At the Computing RRB in April 2009, LHCb presented some resources requests for the 2009 running, in fact for a period covering April 2009 to April 2010. This request was then updated and a final version was approved by the C-RSG in June 2009. As it is not possible to make a report on time for the April RRB on this period we shall compare the requests to the calendar year 2009.

A summary of the June 2009 requests is given by the table below:

Date	Site	kHS06	Disk (TB)	Tape (TB)
Oct'09	CERN	15	720	1000
	Tier-1	31	1740	1000
	Tier-2	31	20	0
Apr'10	CERN	21	1290	1500
	Tier-1	41	3290	1800
	Tier-2	36	20	0
Oct'10	CERN	23	1290	1800
	Tier-1	44	3290	2400
	Tier-2	38	20	0

Table 1: CPU Power, disk and tape storage needed in place to meet LHCb requirements for the 6 month period commencing (a) October 2009, (b) April 2010 and (c) October 2010, as best estimate end of June 2009.

CPU usage in 2009

The CPU figures in table 1 show the actual power that we requested to be able to use, which is higher than the integrated CPU capacity that we need to use in the year, just because the usage pattern is not flat.

The table below shows a comparison between the estimated integrated CPU time and the usage reported by the EGEE accounting portal.

Site	Used (kHS06.years)	Requested (kHS06.years)
CERN	2.18	8.54
Tier1s	8.24	11.7
Tier2s	24.44	17.12
Total	34.86	37.36

Table 2: Comparison between CPU integrated power estimated in the June 2009 request and actually consumed in the calendar year 2009.

One can see that about 90% of the requested capacity was used, but the pattern was different from that estimated at the time. There are several contributing factors to this. The main ones being:

- CERN and Tier1s were barely used for reconstruction and stripping, as there was a negligible amount of real data. Analysis activity was also lower than anticipated due to this. It was anticipated that a fair fraction of simulation would run at CERN and Tier1s until real data arrives. The fraction was lower than anticipated due to high availability of non-Tier1 sites.
- The time interval does not correspond exactly. A lot of simulation was run in the first 3 months of 2009 for tuning the new release of the MC application. This may have increased the Tier2 (read non-Tier1) usage.

At CERN, 32% of the resources used were non-Grid usage, corresponding to analysis, applications tuning, calibration and alignment tests.

Disk usage

Table 2 shows a comparison of disk usage at CERN and Tier1s with the 2009 requests. The resources quoted correspond to end of December unless otherwise stated.

Site	Requested	Allocated	Used
CERN[1] TxD1	650	696.5	482.7
CERN[1] T1D0	70	148.5	irrelevant
CERN[2]	720	721	478
Tier1s[2]	1740[3]	1915	633

Table 3: Comparison between disk resource requirements and disk usage at the end of 2009 (units in TB)

The disk consumption was less than requested, mainly due to the absence of real data. Note that with the first collisions, LHCb collected 217 GB of raw data only, corresponding to about 1 million trigger, out of which about 400,000 were p-p collisions.

---

[1] As measured on Castor on February 15th using Castor tools

[2] As from the WLCG accounting of December 2009

[3] Including 420 TB for T1D0 disk cache in front of MSS

It should be noted that MC data distribution was performed as foreseen in the Computing Model (one copy at CERN, and 2 copies at Tier1s). A clean up of former (so-called DC06) data has been performed except at CERN where data was kept just to allow a few PhD students to complete their work. The whole of DC06 will be deleted before the start of data taking in 2010.

We are not reporting on disk provision nor usage at non-Tier0/1 sites, since the baseline LHCb Computing Model doesn't foresee such a usage, and therefore up to now only storage at Tier0 and Tier1s has been used. We acknowledge the fact that some sites are pledging disk resources, that are currently used either for tests or for providing data access to Tier3 clusters.

We are also not reporting here on the usage of tapes in 2009, firstly as the accounting from sites is not very accurate, secondly because the usage was very moderate due to the absence of real data. A large effort for cleaning up older datasets has been launched in autumn 2009, that should reduce considerably the amount of tapes used at all sites.

Conclusion

Except for what concerns real data that didn't take place as anticipated, LHCb computing activities were pretty close to the expectations. The Computing Model is now fully exercised for simulation, in place for real data, but not tested under stress. This will happen starting in March-April 2010.

LHCb acknowledges the provision by sites of the requested resources. Globally the pledges are satisfactory, although in particular for disk, some sites are still under the required resources while others are providing more than requested, which we appreciate.