

Computing Resources Scrutiny Group

T. Cass (CERN), G.Lamanna (France), D.Espriu (Spain, *Chairman*), J.Flynn (UK), M.Gasthuber (Germany), D.Groep (The Netherlands), D.Lucchesi (Italy), T.Schalk (USA), B.Vinter (Nordic Grid), H.Meinhard (CERN/IT, *Scientific Secretary*)

INTRODUCTION

This report summarizes the deliberations of the CRSG regarding the usage of the computing resources by the four main LHC experiments (ALICE, ATLAS, CMS and LHCb) during 2011.

We have also examined the requests for 2013. Most collaborations have also submitted revised estimates for 2014 but, in view of the uncertainties and diverging estimates, we think it is premature to provide an assessment on the 2014 resource needs. In this report we will deal exclusively with the 2013 request.

Part A of this report is concerned with the overall usage of the WLCG resources and with the scrutiny of the different experimental collaborations' use of these resources, updating the summary presented in the October 2011 C-RRB report. Part B deals with the preliminary review of the 2013 request. A final scrutiny will be provided in the October 2012 C-RRB. The summary of Tier 2 usage kindly compiled by I.Fisk is appended as part C.

2011 has witnessed a full usage of the WLCG resources. The computing models and all the available resources have been subject to real challenges as the amount of data recorded has reached an unprecedented volume due to the excellent performance of the LHC. The planning estimates for 2011 have generally been reflected in reality as the running year has progressed.

We consider the overall performance of the WLCG to be very good and in satisfactory agreement with the estimates and requests. Some exceptions and issues to be improved will be commented on in this report. The accumulated experience makes possible to draw rather firm conclusions and establish some recommendations for the collaborations. These will be discussed below and in section B in connection with each experiment.

The CRSG wishes to express its satisfaction for the outstanding performance of the LHC, of the four experiments under review at this RRB and especially of the WLCG. The LHC has begun to deliver very exciting results and undoubtedly the smooth running of the computing effort has been instrumental in this success.

The LHC running conditions

After the Chamonix meeting early in 2011, it was agreed that the LHC should run for the best part of 2011 (the estimate was 8 months) at a centre-of-mass energy of 7 TeV (3.5 TeV + 3.5 TeV). This schedule has been kept without major disturbances. At the Chamonix meeting held in February 2012 it was decided to increase the energy to 8 TeV (4 TeV + 4 TeV) providing slightly larger cross sections for the processes of interest. About 10% of the time is expected to be dedicated to heavy ion (HI) physics. A long shutdown in 2013 will follow the current run to enable the machine to reach the design energy of 14 TeV (7 TeV + 7 TeV).

Even though 2011 and 2012 were originally envisaged as constituting a single run, in practice the machine has reverted to a yearly cycle.

For the scrutiny the most relevant quantity is the total number of seconds when the beam is declared to be stable and good for physics. Following CERN management recommendations the following scheme has been adopted:

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

Folding in efficiencies $720 \times 0.7 \times 0.4 = 201.6$ effective hours/month = **725760** sec/month

RRB year	RRB year start	RRB year end	Months (max) Data taking	Total live time (in Ms)	pp	AA
2011	April '11	March '12	8	5.9	5.2	0.7
2012	April '12	March '13	8	5.9	5.2	0.7

For the period covered in this usage report presented the estimates have generally been reflected in reality with about 4.7 Ms of pp beams being delivered to the experimental collaborations, or 90% of the theoretical maximum, representing of the order of 1 billion pp recorded events for each experiment (about 10 billion in the case of LHCb). The total live time reflects the excellent performance of the machine and the readiness of the collaborations to take and analyse large amounts of data. The live time fraction is expected to be improved in 2012.

The Pb-Pb run at the end of 2011 was equally successful. In 2012 this run will be replaced by a pPb run of equal length. Unlike in 2011 LHCb plans some activities during this pPb period.

The large number of recorded events has been possible thanks to experiments using all the available bandwidth and effectively recording events at rates larger than the nominal ones. The following are the real/nominal rates for pp events:

ALICE: 380/100 Hz (200 Hz for PbPb, expected in 2012 400Hz, for pPb 560 Hz)

ATLAS: 340/200 Hz (expected in 2012 400 Hz)

CMS: 375/300 Hz (includes 25% data set overlap, expected in 2012 300 Hz)

LHCb: 3000/2000 Hz (expected in 2012: 4500 Hz)

The nominal rates reflect the values indicated in the respective computing models. Some collaborations increased trigger rates after validation of extensions of their respective physics programmes with the LHCC while others have taken advantage of some headroom in CPU capacity and reduction in data sizes to increase their rate. The LHCC in 2011 discouraged a substantial increase in the trigger rates on the grounds that computing resources likely could not be increased in a matching proportion. The CRSG endorsed this recommendation, adding that sustainability of the WLCG in the medium and long run requires a roughly flat budgetary profile in the present circumstances.

Proton bunches are injected with a minimal separation of 50ns rather than the design value of 25ns. To compensate for the reduction in number each bunch contains more protons and they

are squeezed as much as possible to sustain a substantial luminosity (up to 3.5×10^{33} at the end of the 2011 run). The consequence is the appearance of events with many interactions (pile-up). This has a substantial impact both on reconstruction times and on the size of the data sets which are larger than expected because of the increased pile-up with respect to the design conditions and as there is some non-linearity with the number of interactions. Out of time pile-up is also observed.

The rapid growth in peak luminosity came with an increase of the amount of pile-up; in 2011 it was expected to rise up to 15 to 20 events per crossing on average (the largest value refers to the beginning of the fills). However before summer it stayed at a moderate average value of 5-10. After September 2011 the value of β^* was reduced by about 1/3 and pile-up rose to 12-16 events per crossing in average. It is expected to reach the values 24-30 in 2012. At the LHCb IP it was possible to keep the average multiplicity to 1.5 (below the 2.5 peak value for 2010, but still above the design multiplicity).

The collaborations have been forced to revise some of their assumptions regarding data placement policies, number of copies, etc. Collaborations have been diligent in optimising their reprocessing times, reducing raw data sizes and moving towards derived formats for analysis. This optimisation has allowed them not only to cope with the increasing amounts of data generated by the excellent LHC performance and the more complex events due to pile-up, but also allowed them to record at the increased rates indicated before.

During the early months of 2011 the collaborations reconstructed and analysed the events recorded during the first PbPb run. The second PbPb run took place in November 2011 and its result are being analysed now. These runs are particularly relevant for the ALICE collaboration which has revised some computing model assumptions in view of past experience.

Interactions with the experiments

The recommendations of the previous C-RRB report in October 2011 urged the experimental collaborations to submit a detailed account of their use of the resources along with their revised requests by March 1st. CMS and LHCb complied a few days after this deadline while ALICE and ATLAS informed us that they would need some two more weeks to submit their documents. The referees exchanged sets of Q & A with the computing representatives to clarify various points in the respective reports. Several meetings with the collaborations took place too.

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

The CRSG had repeatedly asked the ALICE collaboration to submit requests more aligned with the expected resources and in this way facilitate a realistic scrutiny. Our referees met ALICE representatives over the summer and received assurances that the estimates for 2013 would be realistic and go in the direction of closing the gap between requests and pledges.

Generally speaking the interactions with the experiments are very fluid and we thank the respective managements for their openness and collaboration. Thanks are due in particular to Ian Fisk who compiled and summarized the Tier 1 and Tier 2 usage for 2011.

The CRSG is generally satisfied with the quality and quantity of the information provided by the experimental collaborations. For future reviews we insist that:

- All changes to the models compared to the previous review should be 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 2012 C-RRB meeting this deadline is **September 1st 2012**.

Interactions with the LHCC

Since the last scrutiny no issues appeared which we thought necessary to refer to the LHCC.

Mitigation of the resource growth

Experiments were urged to use the experience gained in the first years of running to modify the computing models in ways that would make them sustainable in the long run and mitigate the growth in computing resources. In the reports submitted to us and in conversations with the collaborations we have seen substantial revisions and adaptations to meet this goal.

- Experiments have made an effort to reduce the raw event size (and the size of all subsequent derived formats) and event processing times. These efforts have mitigated the serious challenge that pile-up represents. This has allowed experiments to record events at a higher rate, indicating some margin of safety and redundancy in the resources.
- Similarly, experiments have set up task forces to reduce processing times and they have generally improved, partly under the pressure to deal with increasing values of pile-up.
- The collaborations have made substantial changes in their data distribution policies, reducing the number of copies stored in Tier 1 or Tier 2 and moved to more compact datasets for analysis. They have been very active in redistributing tasks among CERN, Tier 1 and Tier2 to optimize the usage of resources and an equilibrated distribution of the usage to the point that the boundaries between Tier 1's and some large Tier 2's are disappearing to some extent.
- The collaborations have continued implementing aggressive data cleaning policies
- Substantial progress in the implementation of fast MonteCarlo simulations has been made. Some subdetectors are not amenable to parametric simulations in practice. Generally speaking, however, the experiments have been able to do more MonteCarlo than foreseen.
- The user efficiency is better than planned and continuously improving in spite of user/chaotic analysis being now a visible part of the total but remains an issue in the ALICE collaboration for user/chaotic analysis.
- Some experiments plan to use their HLT farms (or parts thereof) for reprocessing or MonteCarlo production during 2013.
- The collaborations attempt to smoothen out their peak CPU demands throughout the year.

Overall assessment

Generally speaking, the experiments' computing models and the WLCG have demonstrated in a remarkably smooth way their capability to record, distribute and analyse the substantial amounts of data delivered to them by the very successful run of the LHC during last year.

The statistics indicate a massive use of the available resources. Some aspects of the computing models such as large individual non-organized computing usage, format and distribution of the data sets, the flexibility to cope with increasingly challenging running conditions, and the urgency to reprocess and analyse large amounts of data in a short time have represented a real challenge for the computing models and for the WLCG as a whole. This challenge has been passed very successfully.

Noticeable peaks of individual user analysis have been present in the weeks preceding major conferences but the efficiency of the system remained high throughout (with one exception).

CPU resources are generally exceeding the experiments' needs at this point (by a factor that has clearly decreased with respect to previous years) and the experimental collaborations have had substantial headroom that they have employed to increase simulation production. Although less visible, there is still some headroom for disk due to installed resources surpassing pledges in some instances. Computing has never been a limiting factor in any case.

The collaborations have implemented more realistic and more organized data distribution policies. The reprocessing policy is quickly converging to the one indicated in the computing models as the number of events disfavours frequent reprocessing.

The GRID fabric works reasonably well, data distribution and network performance are excellent. A similar comment applies to the middleware.

Some fears that were expressed concerning the availability of disk in the market have not been confirmed. However we insist in limiting the growth on the disk request as much as possible.

Tape is generally underused but this is of little concern since cartridge prices change only slowly and, anyway, most Tier1s will purchase tape as needed to ensure an adequate buffer rather than pre-purchase all of the pledged capacity.

The CRSG welcomes the more realistic approach taken by the ALICE collaboration, adopting a reduced data processing strategy leading to requirements in better agreement with the expected resources. We also welcome their inclusion of resources pledged by non-signatories of the WLCG MoU that are crucial to them to reach their physics goals. We welcome the action taken by some experiments to take into account their non-WLCG resources.

The share of resources among the different Tier 1's, Tier 2's and CERN now looks more even, though one collaboration should still work to reduce the fraction of their total computing done at CERN.

The CRSG congratulates the Tiers and all institutions participating in the WLCG for the overall success of the LHC computing. Special acknowledgement should be given to the personnel and institutes in the many Tier 2 for their essential contribution. A detailed statistics of the Tier 2 usage is appended at the end of this report.

Recommendations

Specific questions related to the experiments' requests are deferred to individual scrutinies

- We recommend the use of the on-line farms during 2013 for reprocessing and simulated data production. This would entail "parking" some fraction of the data for later processing. Thanks to this possibility we envisage a flat profile for CPU in 2013 compared to 2012. Details specific to each experiment will be discussed below.
- We also underline the absolute need to smoothen out the CPU needs throughout the year and consider the possibility of using external (cloud?) resources for very localized demands, particularly for MonteCarlo production.
- The experiments capability to record events at increased rates is very welcome but the CRSG does not see how a substantial increase of the data taking rate could be accommodated with the existing computing resources and does not recommend a formal modification of the computing models in this direction.
- The CRSG would like to keep a balanced usage of the different Tiers. Ensuring such a balance will maintain a healthy WLCG collaboration and so ensure long term success.

- We recommend keeping the request for new disk under close scrutiny. Some collaborations have enlarged their physics scope and this may justify some increases but others have not fully justified the usage of existing disk resources yet. If possible, the collaborations should present data access statistics to better understand and demonstrate that the data placement policies are meaningful and effective.
- Some aspects of the WLCG accounting has improved (e.g. the REBUS database) but there are still deficiencies : the installed CPU compared to the pledged and installed disk capacity at the Tier 2 centres is not centrally accounted so far. It would be useful to disentangle the efficiency of organized/chaotic activities. The CRSG would like to evaluate the assumed efficiency of disk as this is the most expensive commodity. Optimization will be needed to keep costs under control as the number of events keeps increasing. We note that the WLCG master accounting tables still quote the old value for T2 CPU efficiency (60%) rather than the new one (70%). The accounting of T1's at the EGI portal does not seem to be working properly either.
- The CRSG encourages close collaboration of the different centres with the experiments to continue the implementation of intelligent storage management policies to allow efficient and cost-effective access to data. In particular the implications on network bandwidth for best-use of resources should be considered. We consider this issue very relevant for the operation of the LHC experiments after 2014.
- The CRSG recommends a continuous monitoring of CERN's policies for resource sharing when allocations to specific experiments are not fully used.
- We encourage the experimental collaborations to continue working on realistic estimates for the computing needs in 2014 and beyond, keeping the budgetary constraints in mind and working with the CRSG as necessary.

CRSG workplan for 2012

- 1.- The CRSG plans to continue a close follow-up of the implementation of the ALICE computing model.
- 2.- The CRSG would like to follow closely the developments in dynamic data placement policies and have estimates of the eventual impact of these policies on network resources.
- 3.- In order to provide a better scrutiny, the CRSG would like to understand better the efficiency of disk management policies.
- 4- The CRSG strongly recommends a policy of mitigation of the request for new resources by optimizing the use of existing ones as much as possible and shall pay particular attention to this aspect of the scrutiny.
- 5- The CRSG would like to understand better the computing needs after the 2013 shutdown and asks for the experiments collaboration in this respect as this will have an obvious impact on the WLCG budget in years to come.

On the CRSG membership

Concezio Bozzi, representing INFN has been replaced by Donatella Lucchesi. A replacement for William Trischuk (Canada) is now pending. During 2012 it will be necessary to renew or replace those members of the CRSG (including the chairman) that were not replaced during 2010 and 2011.

PART A

Scrutiny of the WLCG resources utilization in 2011

This report refers, unless otherwise stated, to the calendar year 2011, from January 1st to December 31st.

This report has used the following sources:

1.- Cumulative accounting from January to December for Tier 1s and CERN

https://espace.cern.ch/WLCG-document-repository/Accounting/Tier-1/2011/december-11/Master_accounting_summaries_December2011.pdf

2.- Month-by-month accounting of the CPU delivered by the Tier 2s

https://espace.cern.ch/WLCG-document-repository/Accounting/Tier-2/2011/december-11/Tier2_Accounting_Report_December2011.pdf

3.- The EGEE accounting portal at CESGA

<http://www3.egee.cesga.es/accounting/egi.php>

4.- WLCG accounting reports for non-GRID CPU

5.- 2011 pledges as presented to the C-RRB

6.- The Tier-1 and Tier-2 Usage Reports kindly provided by Ian Fisk..

7.- The documents that the experiments have provided to the CRSG.

The following table describes the degree of usage of the different resources. The first set of tables (blue) compile general information, gotten from the WLCG accounting. The set of tables referring to specific experiments (yellow) use information obtained from the collaboration themselves. The latter have been cross-checked with the statistics from the accounting tools whenever possible.

April 2012

Resource	Site(s)	Used/Pledged Period average	Used/Pledged End of period
CPU	CERN	55 %	---
	T1	93 %	---
	T2	166 %	---
Disk	CERN	105 %	119 %
	T1	121 %	137 %
	T2	Not available	Not available
Tape	CERN	75 %	97 %
	T1	47 %	51 %

For comparison we reproduce the analogous table presented in the October 2011 C-RRB that refers to the January-August period

October 2011

Resource	Site(s)	Used/Pledged Period average	Used/Pledged End of period
CPU	CERN	52 %	---
	T1	83 %	---
	T2	117 %	---
Disk	CERN	99 %	99 %
	T1	112 %	116 %
	T2	Not available	Not available
Tape	CERN	64 %	75 %
	T1	47 %	43 %

The CPU figures correspond to a time average over the year obtained from averaging the monthly figures; those for disk or tape are usage relative to the installed capacity at the end of the accounting period.

Efficiencies

The computing TDR estimated 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). This last figure was revised and is now estimated to be 70% for 2013.

The numbers in the case of ALICE merit some comments. While the overall CPU efficiency at Tier 2 is low but still acceptable, the one associated to chaotic analysis drops to a worrisome 16% with a huge dispersion among users. This issue is elaborated further in the ALICE usage report below. Due to the implementation of the ALICE computing model the average efficiency for Tier 1 and Tier 2 is very similar.

LHCb uses T2 mostly for MC production – which is by far more efficient than user analysis.

Efficiency of the utilization of the CPU at Tier 2s per experiment during 2011 (left column) compared to the October 2011 report (right column)

ALICE	54 %	50%
ATLAS	88 %	89%
CMS	83 %	80%
LHCb	93 %	98%

Disk usage

Disk usage is difficult to analyse. A metric based exclusively on disk occupancy does not account for frequency of access or how efficiently disks are managed.

From the available information the CRSG has been unable to verify the theoretical efficiency of disk, which is a very relevant ingredient of the total cost. We encourage the larger experiments to provide more information to the CRSG in this respect.

Sharing of the WLCG resources

The following tables give an idea of the use by 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). On the first (CERN+Tier 1) table 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%). For comparison the percentages reported in October 2011 are shown in a separate table.

Percentage of use of the resources by experiment in 2011 (CERN+Tier 1s)

Collaboration	% of tape inT1+CERN used at end of period	% of disk inT1+CERN used at end of period	% of CPU in T1+CERN used	% of which at CERN
ALICE	12 %	14 %	15 %	52 %
ATLAS	39 %	46 %	51 %	17 %
CMS	41 %	33 %	23 %	21 %
LHCb	8 %	7 %	11 %	26 %

Percentage of use of the resources by experiment in January-July 2011 (CERN+Tier 1s)

Collaboration	% of tape inT1+CERN used at end of period	% of disk inT1+CERN used at end of period	% of CPU in T1+CERN used	% of which at CERN
ALICE	11 %	13 %	13 %	59 %
ATLAS	41 %	47 %	52 %	18 %
CMS	42 %	32 %	23 %	18 %
LHCb	7 %	8 %	12 %	28 %

The metrics show that the trend described in the October 2011 report has been maintained. ALICE has decreased slightly their dependence on CERN resources. The figure for LHCb has been maintained at a moderate level, in contrast to previous years. Both are welcome developments.

Percentage of use of the resources by experiment in 2011 (Tier 2s)

Collaboration	% of CPU in T2 used (All 2011)	% of CPU in T2 used (October 2011)
ALICE	11 %	6 %
ATLAS	54 %	54 %
CMS	33 %	31 %
LHCb	2.5 %	4 %

We note the increase of T2 CPU usage by the ALICE collaboration in the last months of 2011. Other variations are not significant.

Delivered versus pledged

The overall level of fulfilment of the pledges can be seen from the following table.

Resource	Site(s)	Installed / pledged
CPU	CERN	100 %
	T1	99 %
	T2	136 %*
Disk	CERN	100 %
	T1	109 %
	T2	Not available
Tape	CERN	100 %
	T1	89 %

The figures refer in all cases to the end of the reporting period. There have been noticeable shortages at NL-LHC-T1 and NDGF with respect to the pledge in CPU (75% and 77%, respectively) and disk (88% and 81%, respectively). Tier1 have typically bought less tape than planned reinforcing the view that this is an underused resource (usage was 51 %). The Tier 2

CPU(*) percentage quoted is delivered/pledged as there. At the end of 2011 the installed/pledged ratio referred to 2011 was 188 %. Comparing the 109% of installed disk@ T1 (with respect to the pledge) with the 121% used shows that the efficiency surpasses the theoretical 70%. Additional resources are clearly visible in several sites.

Usage by the individual experimental collaborations

In what follows CPU usage refers to the average over the period. Disk and tape usage refers to the occupancy at the end of period. Units are kHS06 and PB for CPU and memory, respectively. Data are provided by the experiments and cross-checked with the WLCG accounting tools whenever possible. Numbers do not always coincide.

ALICE usage

Resource	Site(s)	2011 request	2011 pledge	2011 usage	Efficiency
CPU/kHS06	T0+CAF	62	62	56	67%
	T1	117	71	60	59%
	T2	121	81	107	54%
Disk/PB	T0+CAF	6.1	6.1	5.0	--
	T1	7.9	5.5	6	--
	T2	6.6	7.3	9.9	--
Tape/PB	T0+CAF	6.8	6.8	7.9	--
	T1	13.0	8.0	3.5	--

Comments on the ALICE usage report

The ALICE resources requested for 2011 were well above the pledges made for Tier 0 and Tier 1, but close for Tier 2. The usage figures (from various sources) provided by ALICE show that CPU use is close (below) the pledge at all Tiers. Disk usage on the contrary is above the pledges indicating that ALICE has alleviated some of their chronic shortages taking advantage of opportunistic disk use. Tape use at CERN is above the pledge, while at Tier 1 it is considerably below. About a quarter of CPU and disk resources used in 2011 were provided by CERN (23% of CPU wall time and 24% of disk storage, a fraction that is proving impossible to sustain). The most obvious issue in usage has been CPU efficiency.

CPU efficiency was poor at all tiers near the start of the 2011 RRB year. This was traced to a combination of the way codes accessed the Offline Conditions Database (OCDB) and a high proportion of user analysis. The OCDB is now accessed more effectively so that simulation and scheduled analysis run efficiently but the efficiency of user analysis remains very variable and the collaboration continues to address this serious problem. However, ALICE pointed out to us as possible reasons that disk space is saturated at some sites resulting in user jobs writing data

remotely, and/or users submit development/test jobs to the grid rather than using local resources. However, the CRSG is of the opinion that a more structured model could increase the efficiency.

Analysis facilities (AFs): ALICE has three common AFs at CERN, Lyon and in Slovakia. Two more, in Nantes and Dubna are used exclusively by local physicists, while a fourth common cluster in South Korea is in preparation. The AFs are used during data-taking to reconstruct and analyse a fraction of events for quality control and for data calibration. They hold subsets of RAW and MC data to be analysed for early physics discovery and are used for testing, tuning and improving analyses. We note this and encourage ALICE to use the AF as much as possible for development work, avoiding the most inefficient ALICE user jobs running on the grid.

Disk usage: ALICE has launched a campaign to reduce storage. Old data and MC from before 2010 is removed; data not heavily used is kept in only one copy (but current data still has multiple copies). The collaboration is trying to move from ESD to AOD for user analysis and will then keep only one ESD, but several copies of the smaller AOD.

Raw data storage: All raw data is kept in mass storage at CERN but, to reduce tape requirements at Tier 1, only data which passes quality checks for physics analysis is distributed for storage at the Tier 1 sites. The trigger mix used in 2011, with a larger fraction of smaller MUON-triggered events, reduced the average event size, further reducing Tier 1 tape usage.

ATLAS Usage

Resource	Site(s)	Pledged	Used	Used/ Pledged	Average CPU efficiency
CPU (kHS06)	T0+CAF	75	82	109 %	90 %
	T1	248	244	99 %	87 %
	T2	285	405	142 %	88 %
Disk (PB)	T0+CAF	7	5	70 %	-
	T1	26	27	103 %	-
	T2	35	22	62 %	-
Tape (PB)	T0+CAF	12	14	117 %	-
	T1	32	16	50 %	-

Comments on the ATLAS usage report

The status of the ATLAS offline computing, the resource usage and the 2012-2013 requests were reported by the experiment to the CRSG on 24 March 2012. ATLAS provided a detailed report on the computing resource usage in 2011, together with a comparison with the predicted needs.

ATLAS recorded data at a rate of 340 Hz (nominal rate is 200 Hz) which increased above 500Hz when the calibration triggers are included. Data compression worked for RAW, ESD, AOD and many simulation data sets. Reported events sizes were 50% of previous data. Full simulation processing time was reduced by 50% and the real reconstruction time by 30%.

ATLAS is continuing to move their physics analysis model toward using derived data (D3PD) which are Root N-tuple files small enough to fit on local resources. The T2 disk usage continues to be substantially underused. The T1 tape usage was 55% of predicted.

They have processed roughly 4.5 times more simulated data than originally planned, thus indicating that CPU and disk resources are more than sufficient at the present stage. They have also begun to use the fast MC simulation for some specific analysis.

The implementation of the ATLAS computing model seems satisfactory. We note however the low degree of usage of disk at the Tier 2 and tape at the Tier 1, two issues that are recurrent in our reports.

CMS usage

Resource	Site(s)	Pledged	Used	Used/ Pledged	Average CPU efficiency
CPU (kHS06)	T0+CAF	106	39	37%	59%
	T1	130	114	88%	85%
	T2	305	265	87%	80%
Disk (PB)	T0+CAF	5.4	3.7	68%	-
	T1	16.2	15.5	97%	-
	T2	18.1	12.7	70%	-
Tape (PB)	T0+CAF	21.6	14	65%	-
	T1	45	30	67%	-

Comments on the CMS usage report

The performance of the CMS computing system was generally smooth throughout the whole period.

CMS gave their Resource Utilization document and Resource Request to the CRSG at the beginning of March. They showed their performance numbers for the 300Hz trigger (375 HZ if 25% overlap is accounted for) taken during 2011 running. During this period the LHC performance averaged to about what was planned and CMS software performed close to predictions with the exception of reconstruction time and memory footprint. At the end of 2011 the observed pile-up was on average between 16 and 17 interactions per crossing.

CPU utilization at the T0 was below 70% due to the large memory footprint. Unfortunately success in reducing memory demands by 30% has been somewhat offset by the recent increase in pile-up.

CMS has successfully made the transition to a 64 bit code base and a new and improved application I/O layer. CMS recently moved to a disk only storage model for the CAF to improve the low analysis CPU utilization.

The model CMS is using to evaluate the effect of the pile-up has been tested in 2011. The event size generally agrees with the expectations for low pile-up. The event reconstruction time in 2011 has been 20% higher than the planning estimates for the same pile-up conditions.

In 2011 CMS has reduced reprocessing to the number of passes foreseen in the computing model; no additional resource load is expected for this activity.

Their Tier1 efficiency is high (87%) and the Tier2s are running with efficiency above 80%. They have successfully made a transition to AOD usage for their analysis and will only produce AOD files from most of their simulation production to save space.

During 2011 the T1 centres did more simulation work to free up the T2's for analysis and did a complete reprocessing of the data at the end of calendar year 2011. The total number of digitized and reconstructed simulation events is significantly ahead of projections. There are 2.8B new events simulated in 2011 and 1.7B events reprocessed with new pile-up and reconstruction codes.

The computing details are close to expectation from the computing model with a few exceptions.

LHCb usage

Resource	Site(s)	Pledged	Used	Used/ Pledged
CPU (kHS06)	T0 (CERN)	21.0	7.2	34 %
	T1	69.2	40.0	58 %
	T2 + others	40.5	47.0 + 26.2	180 %
Disk (PB)	T0	1.5	1.2	80 %
	T1	3.7	2.7	73 %
	T2	--	--	--
Tape (PB)	T0+CAF	2.5	2.1	84 %
	T1	3.9	3.3	85 %

Comments on the LHCb usage report

The usage of the computing resources by LHCb displays a rather healthy situation. The experiment was successful during 2011 in using the resources and adapting to the contingent situation of higher interactions rate and a broadened physics program, although without some difficulties.

During 2011 LHCb required resources larger than the already known pledges. The LHCb experiment indicated that Tier0/Tier1s could not provide the peak of CPU power needed to reprocess the full 2011 data sample in time for the winter conferences and some extra resources from Tier2s had to be used. However the usage of Tier0/Tier1s is on average at the level of ~50% of the pledges. Opportunistic use of Tier2s was also critical for the 2011 MC production which required about twice the amount of the available CPU. Both effects approximately compensate and LHCb has managed to keep within the available pledges overall thanks to the flexibility of the tools and the computing model. However, the previous considerations suggest that the LHCb computing model requires some rethinking.

The use of storage resources is the part of the LHCb computing model where more changes have taken place. The 2011 re-assessment revealed that disk, mainly at Tier 1s, would not have been enough to store the increased data samples without substantial changes. Previous estimates were based on an incomplete consideration of realistic LHC parameters and did not take into account the new charm data.

The CRSG appreciates the efforts made by the LHCb collaboration (in fruitful cooperation with all Tier0/1s) to follow up the CRSG recommendations by adopting a severe policy of cleaning up of disk space and reducing replicas (of older version of data samples) and space

A simulation tool was adopted to redesign the LHCb computing model and re-evaluate the 2012 needs taking into account data-taking conditions, beauty and charm trigger rates, data sizes and volume, data reconstruction rates and CPU requirements, rate of simulated events, safe number of archive and disk replicas, and a modeling of users' activity. The expansion of the LHCb physics program requires additional CPU and storage resources and during the first part of 2012 LHCb had to make substantial efforts to fit within the pledge at the cost of delaying the physics results. By making adjustments (and using non-pledged resources) the collaboration is confident to be able to meet its goal during 2012.

PART B

Scrutiny of the requests for 2013

ALICE

The collaboration has reviewed its computing request to bring it more into line with 2012 pledges and the expected future availability of resources. The requirements for 2013 are given in the table below.

In 2012 ALICE will benefit from additional non-WLCG resources to bridge the gap between the pledges and the requirements. It is expected that some of these resources will become part of the WLCG during 2012 and that ALICE will then gain two new Tier 1 centres, in South Korea and Mexico. Discussion is ongoing in Russia to provide Tier 1 service from 2014–15 and there is a possibility of an Indian Tier 1 centre on the same time scale.

We note some effects of the LHC schedule on the ALICE requirements:

- The heavy-ion run will be protons colliding with lead ions (p-Pb), with the experiment planning to collect 3×10^8 events. The collaboration will collect a bigger fraction of (larger) minimum bias events of this new type and thus the requirements assume the same data sizes and processing times as for Pb-Pb events.
- For pp running at 4 TeV, ALICE will arrange its data taking and trigger setup to collect sufficient reference data at the new energy. ALICE aims to collect 1.4×10^9 pp events and the average event size used in the requirements takes into account the trigger mix.
- With no LHC running in 2013, ALICE will perform first pass reconstruction of one third of the 2012 p-Pb events in the 2012 RRB year and reconstruct the remaining in 2013.

- The long shutdown will also be used to process a backlog from 2011 and 2012.

ALICE has reduced raw event sizes by implementing “HLT compression”. TPC clusters are calculated on-the-fly and the cluster is saved rather than the pad hits. This reduces the raw data size for Pb-Pb events by about a factor of 4. In consequence, more events can be saved, tending to increase subsequent CPU demand. However, there has also been a large reduction in the average CPU consumption per event for heavy-ion reconstruction (this change is a consequence of the mix of events processed rather than a reduction for individual event types). Both of these changes are incorporated in the 2012 and 2013 requests.

The 2012 CPU requirement was reduced by scaling the amount of analysis and MC simulation. This adjustment procedure is carried forward for 2013 so that the requested CPU power does not increase.

The collaboration aims to match the numbers of MC events with real events. To reduce the demands of heavy-ion MC, base events are calculated and reused 10 times with signal overlaid. We asked about parametric or fast Monte Carlo. Parameterisation has been done for calorimetry, but ALICE has not been able to parameterise the TPC tracking which accounts for 80% or so of the MC simulation time. However, ALICE has developed a method of embedding MC signal into real heavy-ion events which reduces CPU needs for heavy-ion simulation.

ALICE is not planning to use its online farm for offline computing during the long shutdown. To do so would necessitate substantial reconfiguration. In addition, the farm will be out of action for much of the shutdown for hardware upgrades.

As noted in the usage report for 2011, ALICE is working to reduce disk storage by more ruthlessly pruning older data and moving towards smaller derived data formats for analysis. Nonetheless, disk space is an issue for the collaboration and leads to low CPU efficiency, when jobs must read or write data from or to remote locations.

In summary, we acknowledge and commend ALICE’s efforts to moderate its requests in light of the expected availability of resources and thank the collaboration for reporting non-WLCG resources, showing how they help reduce the gap to requirements. CPU efficiency remains below that of the other LHC experiments. While we acknowledge ongoing measures to improve this and the effect of lack of disk space, we nevertheless strongly encourage ALICE to take advantage of the long shutdown to make further efforts to match the CPU efficiencies of other experiments.

Resource	Site(s)	2013
CPU/kHS06	T0+CAF	125
	T1	95
	T2	195
Disk/PB	T0+CAF	13.4
	T1	10.9
	T2	19.4
Tape/PB	T0+CAF	23.5
	T1	19.1

ATLAS

ATLAS plans to make several changes to their computing resource utilization in 2012 based on their 2011 experience. Their resource requirements are based on doubling the real trigger rate (to 400 Hz) and will log but not prompt reconstruct an additional 130Hz of special triggers. These delayed streams will be reconstructed in 2013 as part of a general reprocessing of all data taken to that point and a re-do of corresponding simulation.

The events will be larger and take more time to process due to the mean number of interactions per beam crossing being as large as 25. Major reprocessing and larger pileup imply a significant increase in resource usage (both CPU and disk). Event sizes for real ESD, real AOD, simulated ESD and simulated AOD are expected to double in the 2012 running.

In their new model ATLAS will change the details of their data samples kept on disk relying on their "PanDA Dynamic Data Placement" (PD2P) mechanism which they have deployed to all ATLAS clouds and also on a data compression scheme introduced in July of 2011.

Given their experience of a larger than projected simulation production in 2011 they plan to generate a similar volume of simulation data in 2012 and to re-do of all simulation corresponding to 2010-2012 data in 2013. ATLAS anticipates being able to use their High Level Trigger farm during 2013 for some of this simulation production. During 2011 the full GEANT4 simulation time for an event was reduced by a factor of 2 and work is ongoing to reduce it even further.

ATLAS submitted a 'revised' 2012 request. The resources for 2012 are already in place so this is a purely theoretical exercise. However, the CRSG has found these estimates useful to make its assessment for 2013. 2012 will be very much similar to 2011 as data taking is concerned, except that pile-up will increase noticeably. Experiments will not take more data in 2013 and they can use all the resources for analysis and MonteCarlo production. In addition at least a part of the HLT farms will be available. Taking into account the LHCC recommendations and having the previous considerations in mind we conclude that the committed resources should match the revised 2012 ones.

We remind the collaborations that, while they are welcome to write data at increased rates they cannot expect that resources automatically increase to match these rates. Therefore they should be selective in the kind of 'dark' or 'parked' data they plan to collect. Maintaining a reasonably flat profile is essential for the sustainability of the WLCG. We note that this is a tentative scrutiny; the final one will be provided in the October 2012 C-RRB where the present estimates can be revised if deemed necessary.

The approved requests are given in the following table. The tentative estimate given in October 2011 is also provided for comparison

CPU [kHS06]	2013 (this scrutiny)	2013 (previous estimate)
CERN	111	111
Tier-1	297	273
Tier-2	319	281
Disk [PB]		
CERN	10	10
Tier-1	29	30
Tier-2	49	53
Tape [PB]		
CERN	19	18
Tier-1	34	33

CMS

CMS based their request for 2012/2013 computing resources on live time of 5.2 Ms, higher energy, pileup of 30 and an increased trigger rate. They plan to log data for prompt reconstruction at 300 Hz and additional “parked data” stream at an average of 400 Hz. This parked data will be reconstructed at the T1’s, 50% in 2012 and the rest in 2013 after the LHC beams are turned off. The tape needed for this extra parked data stream has been made to fit within the original request for tape.

Overall their entire total computing resource request is unchanged in 2012 but they have changed their computing model for 2013 to move all the CAF resources into the T0 and to shift their reprocessing cycle, normally done at the end of each year, into 2013 to allow the T1’s to reconstruct ½ of the parked data in 2012. The 2013 request for CPU has increased by 15-20%.

They have made progress on their reconstruction time and memory usage, speeding up the reconstruction time by a factor of 2.5 and reducing the memory requirement by ~ 30%.

We remind the collaborations that, while they are welcome to write data at increased rates they cannot expect that resources automatically increase to match these rates. Therefore they should be selective in the kind of ‘dark’ or ‘parked’ data they plan to collect. Maintaining a reasonably flat profile is essential for the sustainability of the WLCG.

CMS also submitted a ‘revised’ 2012 request. The resources for 2012 are already in place so this is a purely theoretical exercise. However, the CRSG has found these estimates useful to make its assessment for 2013. 2012 will be very much similar to 2011 as data taking is concerned, except that pile-up will increase noticeably. Experiments will not take more data in 2013 and they can use all the resources for analysis and MonteCarlo production. In addition at least a part of the HLT farms will be available. Taking into account the LHCC recommendations and having the previous considerations in mind we conclude that the committed resources should match the revised 2012 ones.

We note that this is a tentative scrutiny, the final one will be provided in the October 2012 C-RRB where the present estimates can be revised if deemed necessary

	2013	2013
CPU [kHS06]	(this scrutiny)	(previous estimate)
CERN	121	120
Tier-1	145	145
Tier-2	350	306
Disk [PB]		
CERN	7	7
Tier-1	26	27
Tier-2	26	26
Tape [PB]		
CERN (including HI)	23	23
Tier-1	45	59

LHCb

The LHCb computing resource request for 2013 is based on the experience of last years and a newly introduced simulation model. LHCb takes into account the new charm physics extended reach and includes modifications of the computing model in order to stay within the resource boundaries of the previous year. As outlined in previous reports, 2013 will be a year with no data taking.

The assumed running time is 5.0×10^6 seconds for 2012 assuming an overall LHC duty cycle of 30%. According to the new LHC schedule, data taking is planned for 2012 similar to 2011 with slightly changed parameters. The running time estimates used by LHCb are based on the agreed machine running scenarios. Contrary to the other LHC experiments, LHCb will not take data during the heavy ion runs but may participate for testing and scaling purposes not resulting in any resource requests.

The overall situation for 2013 is affected by the following unusual conditions

- the amount of data taken in 2012 (due to the extended physics reach of LHCb) does not fit to the pledged resources for that year. Part of this data get 'locked' in 2012 and 'unlocked' in 2013 resulting in normal reco and analysis tasks as seen for data taking periods also for 2013.
- the changed LHC parameters requiring more MC production during this transition phase.

The request of computing resources is summarized in the following table

Site	kHS06	Disk (PB)	Tape (PB)
Tier-0	21	3.5	6.2
Tier-1	55	7.6	6.1
Tier-2	47	0	0
Unpledged	(54)	--	--
Total	123 (177)	11.1	12.3

Next table shows the peak usage broken into tasks and the maximum peak expected throughout the year (not the sum of the above). Thanks to the flexibility of the software nearly all of the tasks listed in table 2 can now run in principle at any tier – clearly reflecting a change to the classical Monarc model with reduced barriers between the different tiers

Task	2012	2013
MC	101	101
Physics	49	49
Reco	51	0
Repro	89	109
Total	236	255

LHCb was able to demonstrate the usability of the farm resources for certain tasks and plans to use it during 2013 but no primary usage for this resource is outlined yet. The use of the on-line farm will ease the demand on other resources.

The CRSG thinks that the resources requests for 2013 are based on realistic estimates. The extended physics reach, resulting in even higher trigger rates (now 4.5 kHz for 2012), theoretically demanding more CPU and storage resources, will not lead to higher requests because of several compensating changes to the computing model that have been described to the CRSG..

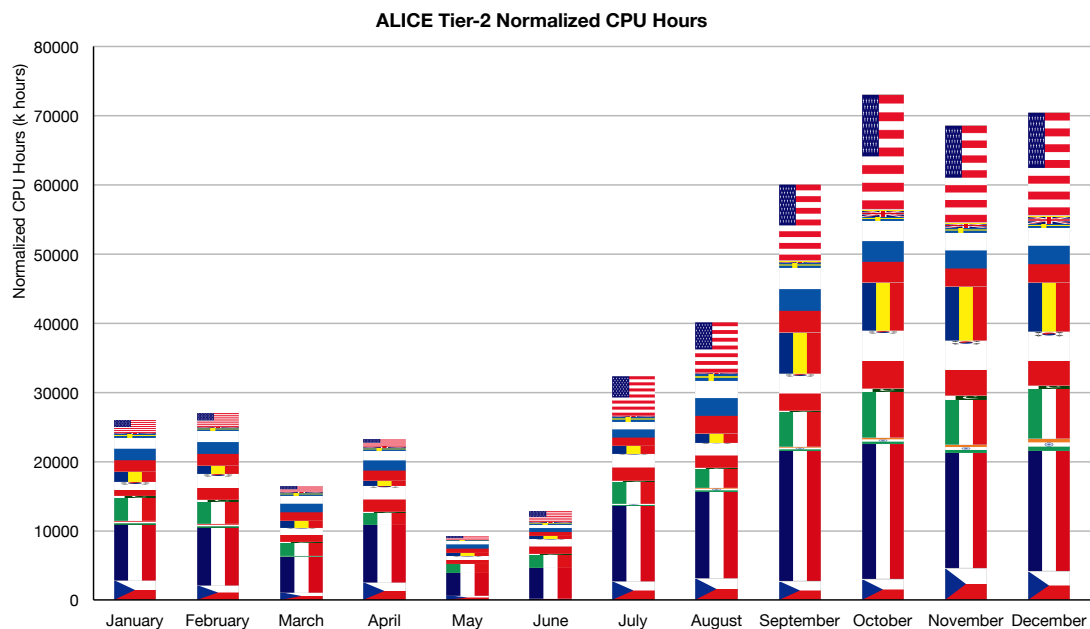
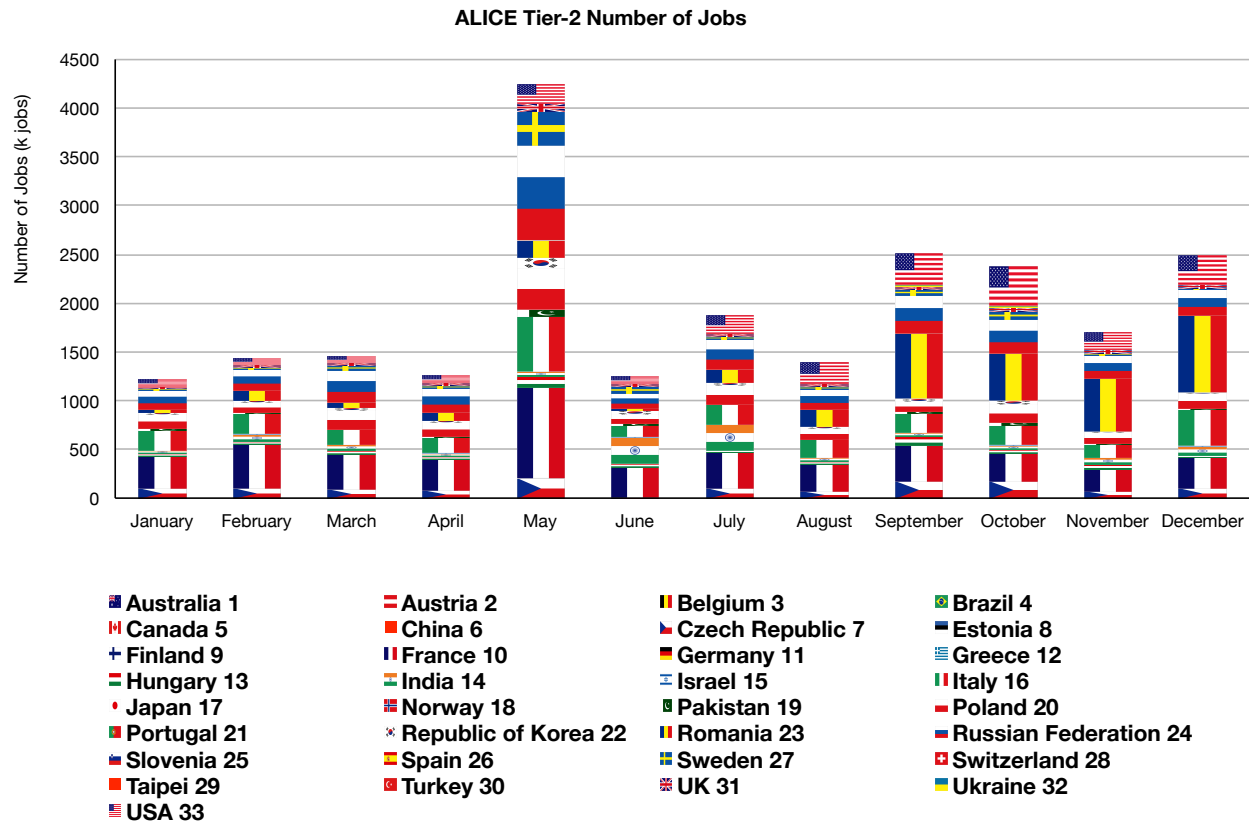
The referees support the ability of LHCb to efficiently use any CPU resource available, mostly independent of the Tier-Level. This should help to cope with high peak CPU usages and at the same time utilize idle resources for time uncritical MC production. The usage of unpledged resources is welcome but clearly these resources are not totally guaranteed and planning should be done carefully.

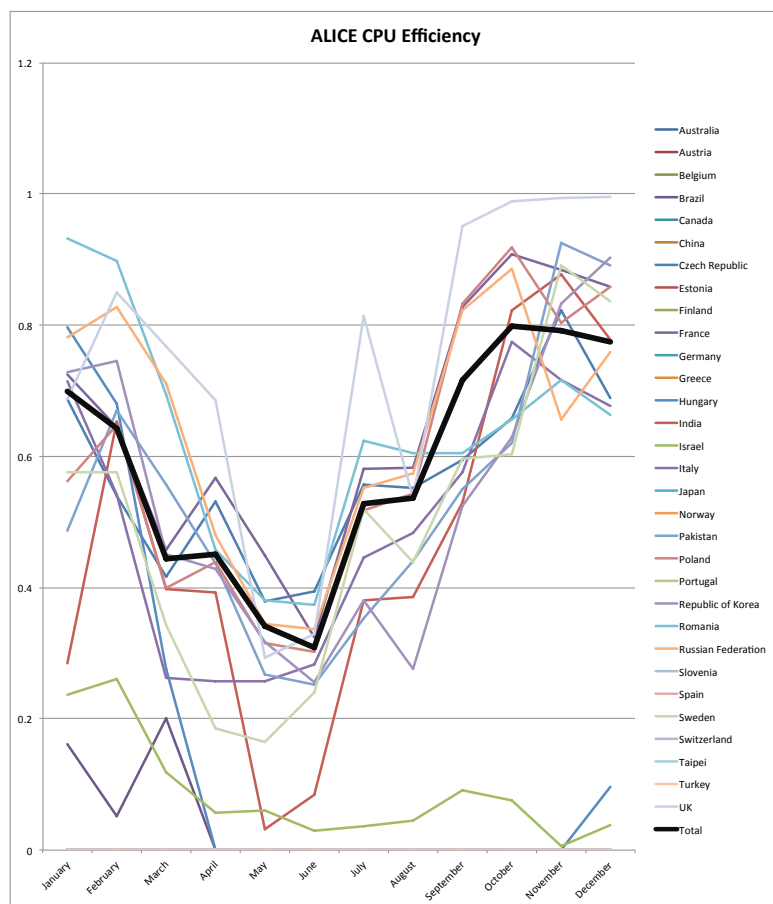
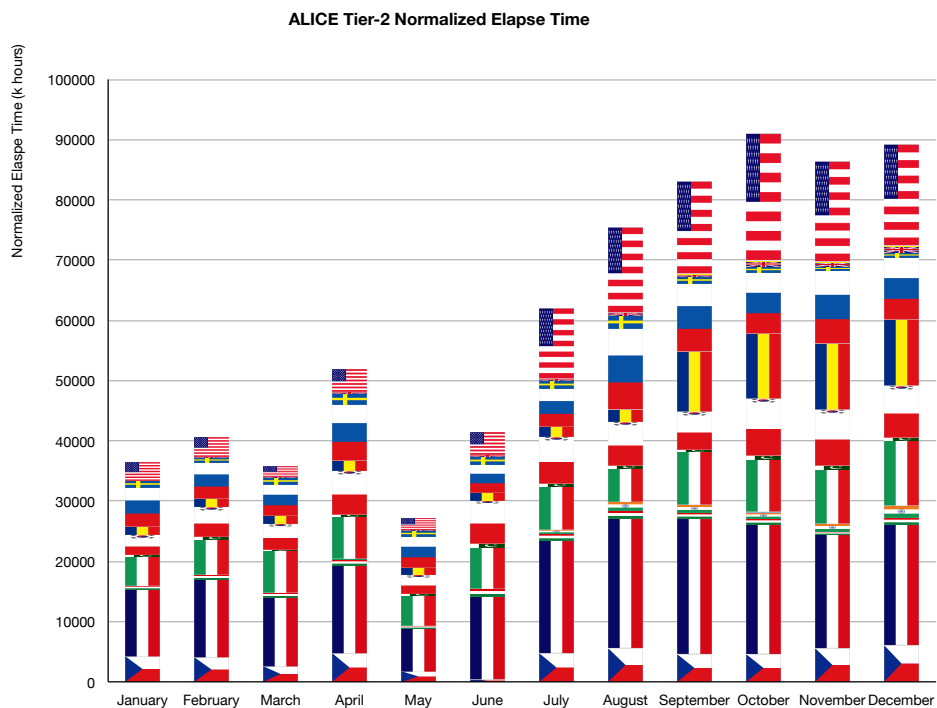
PART C

Usage of the Tier 2's

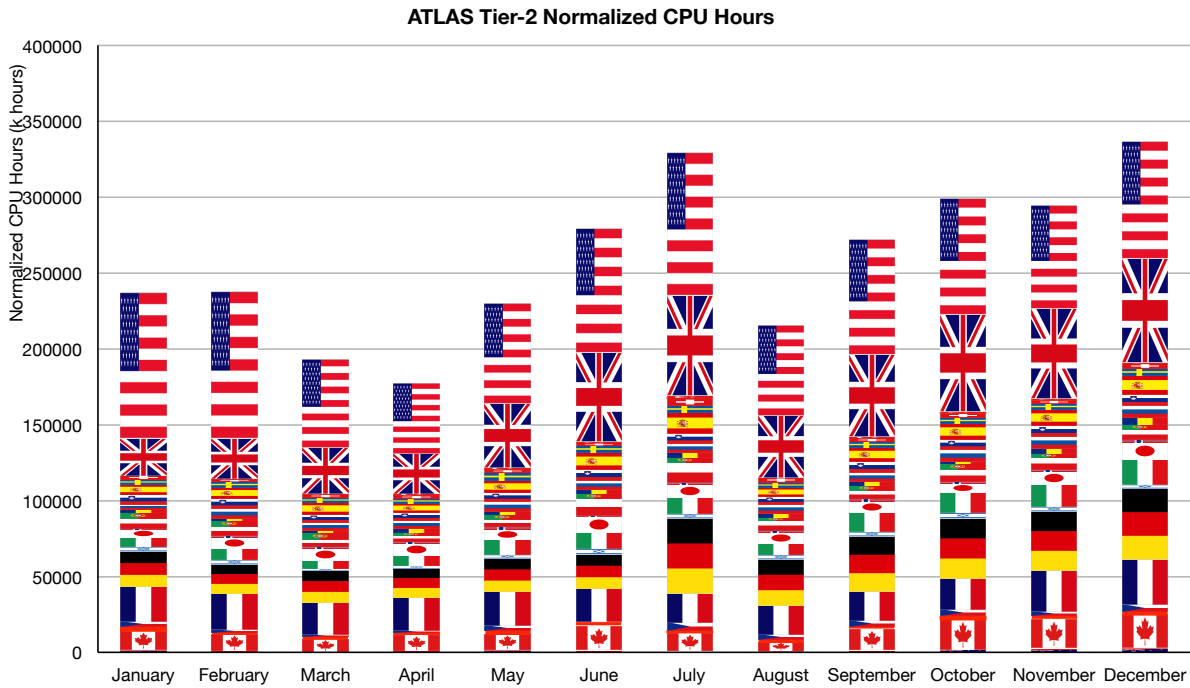
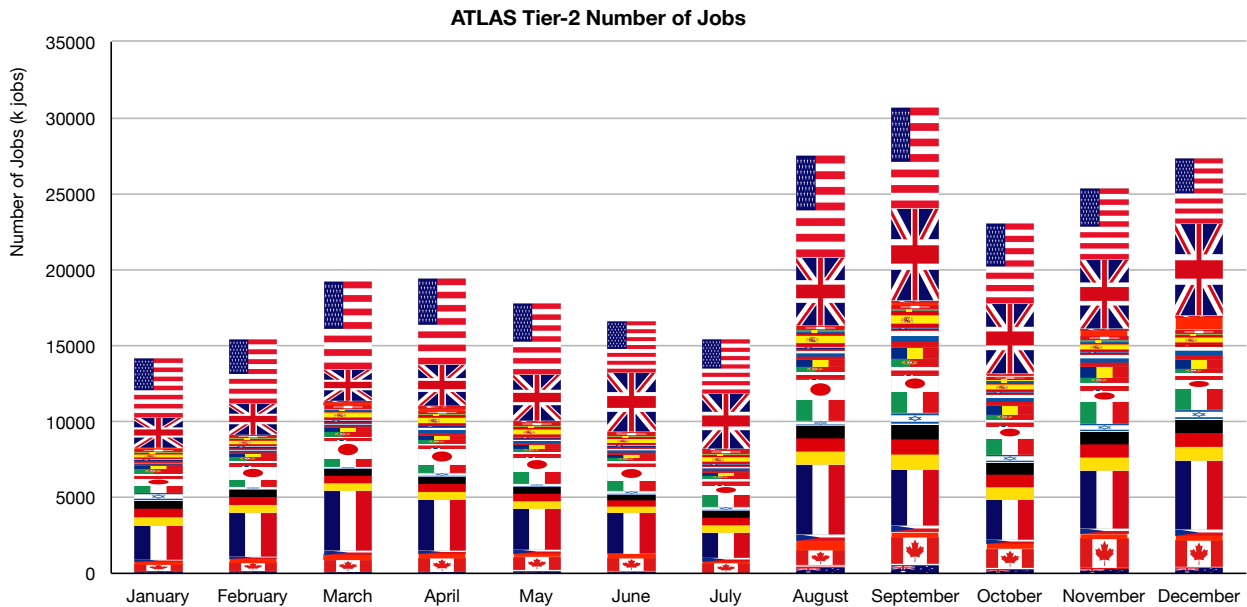
Tier-2 Experiment Reports

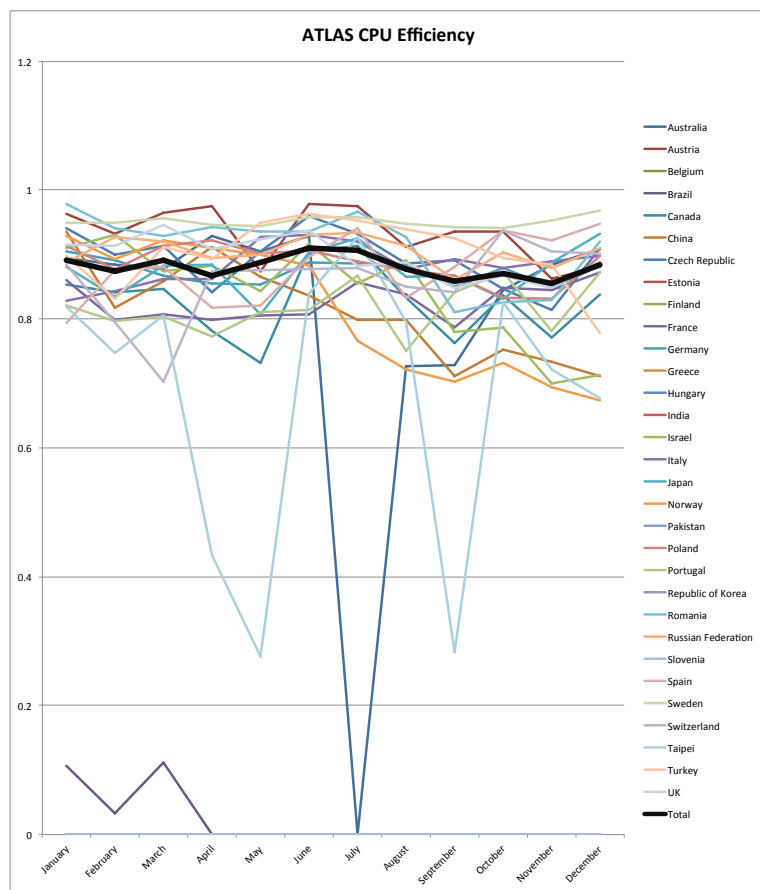
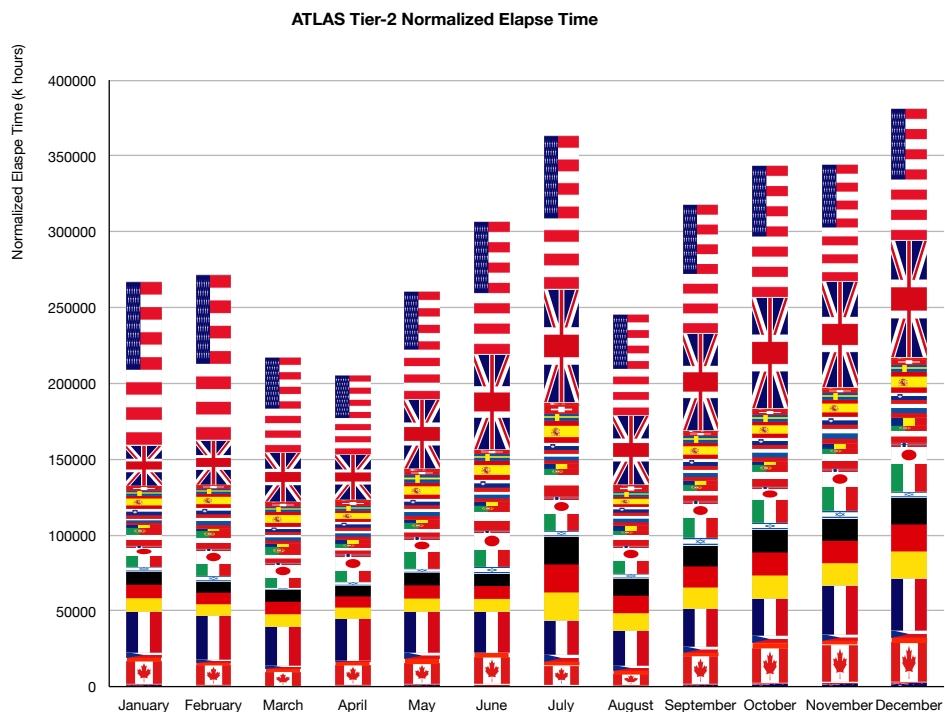
ALICE Plots 2011:



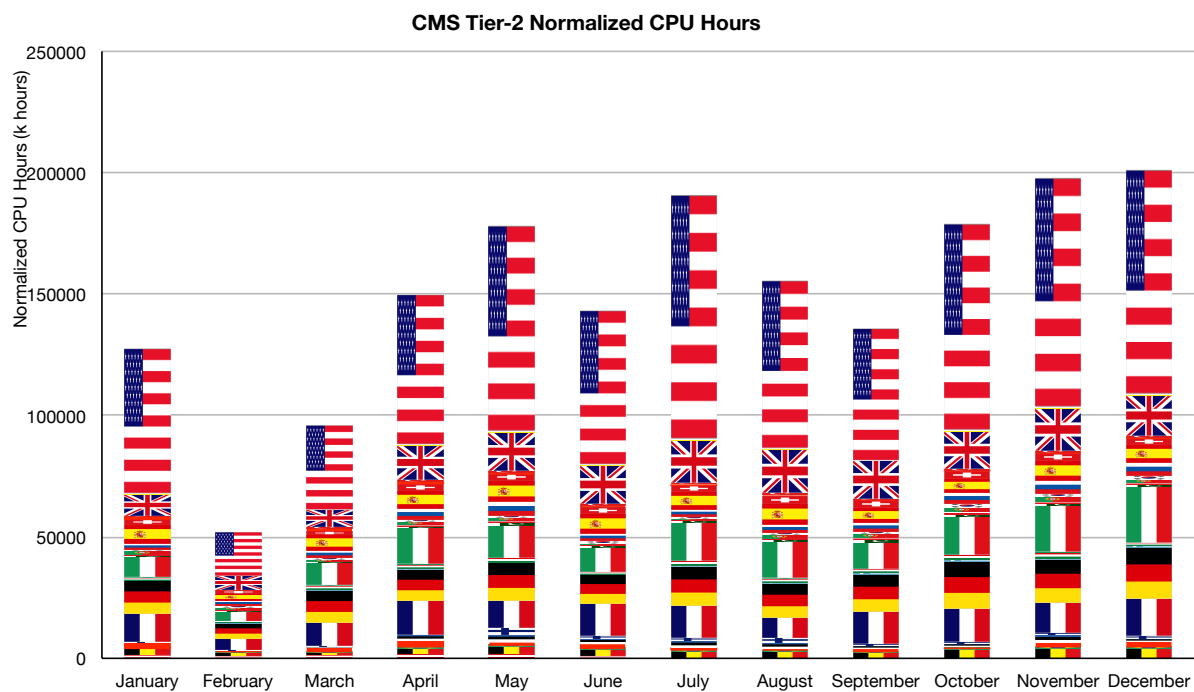
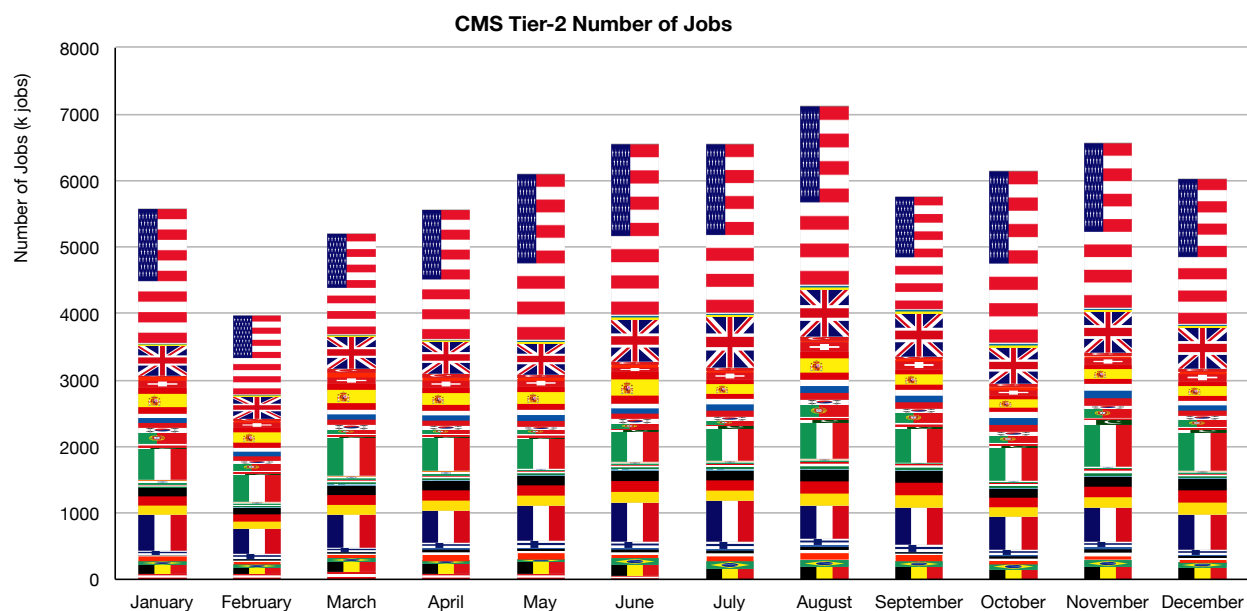


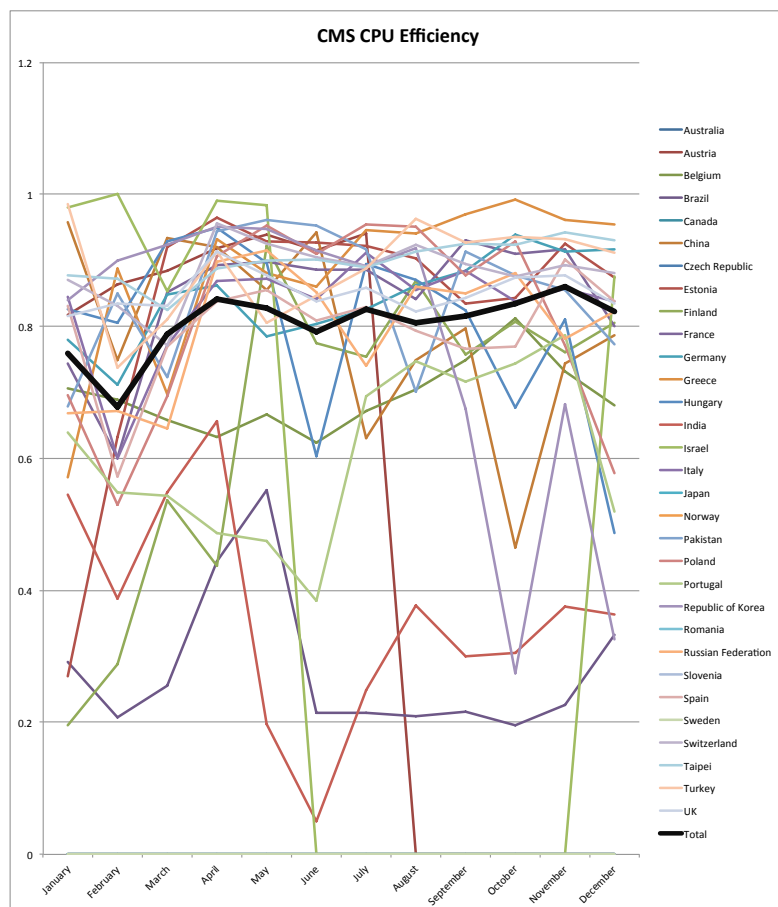
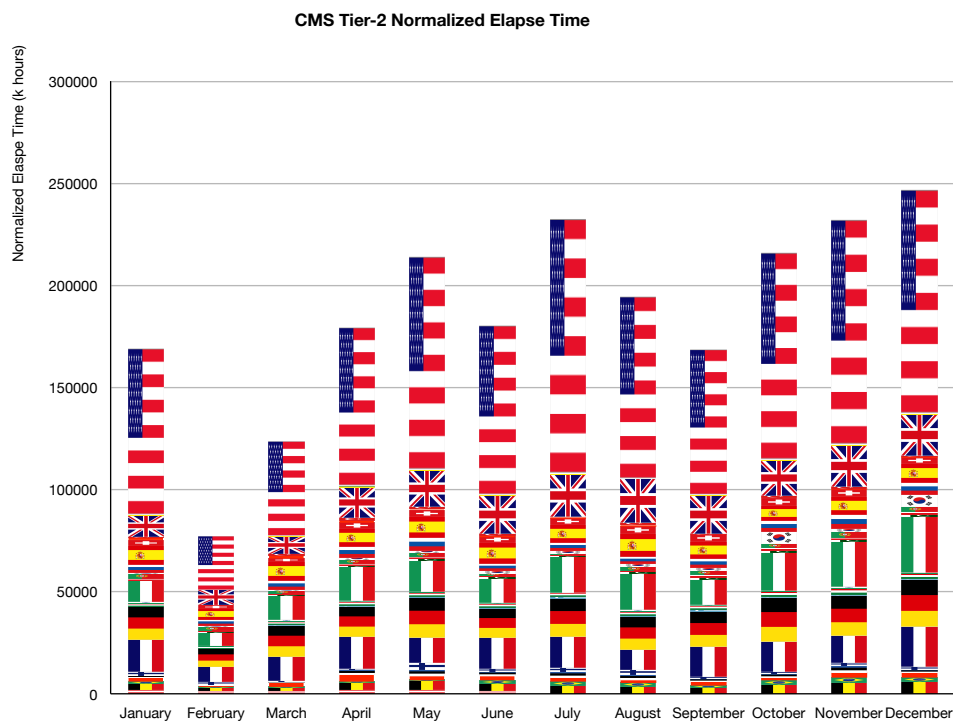
ATLAS Plots 2011:



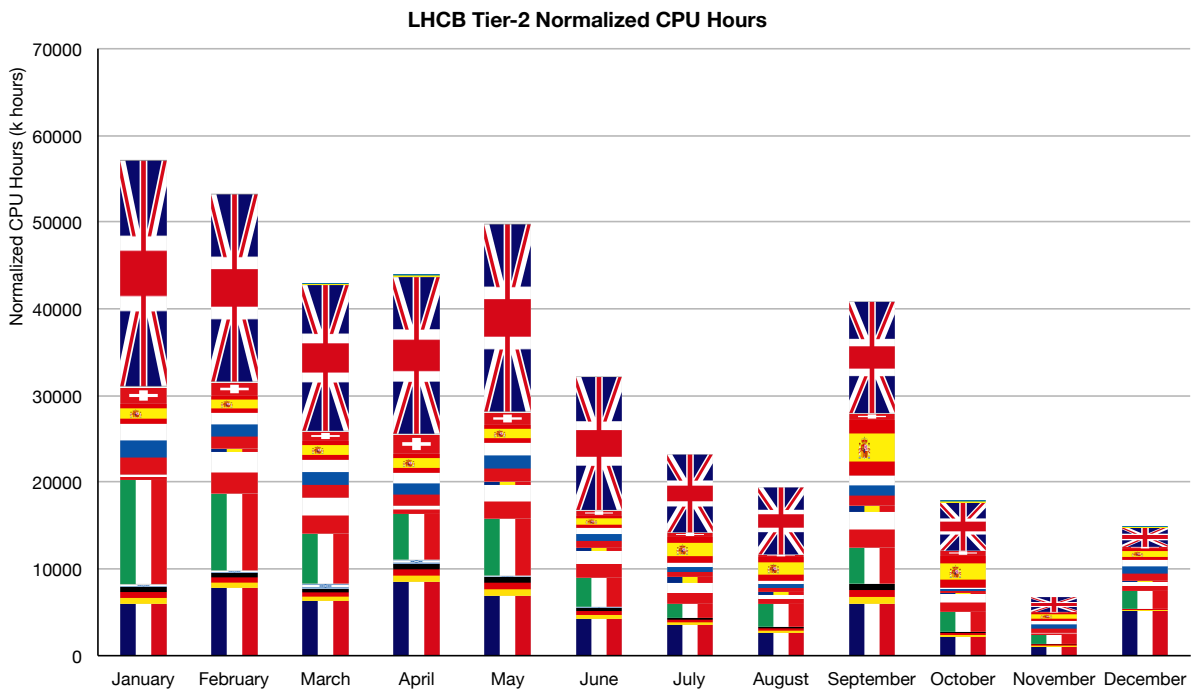
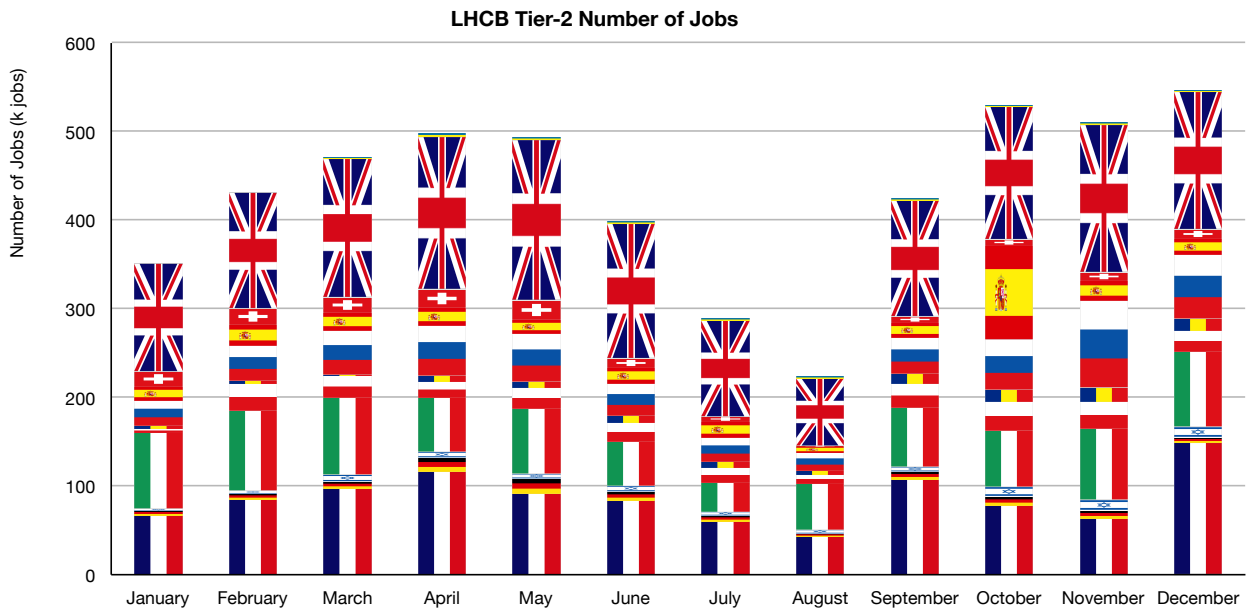


CMS Plots 2011:

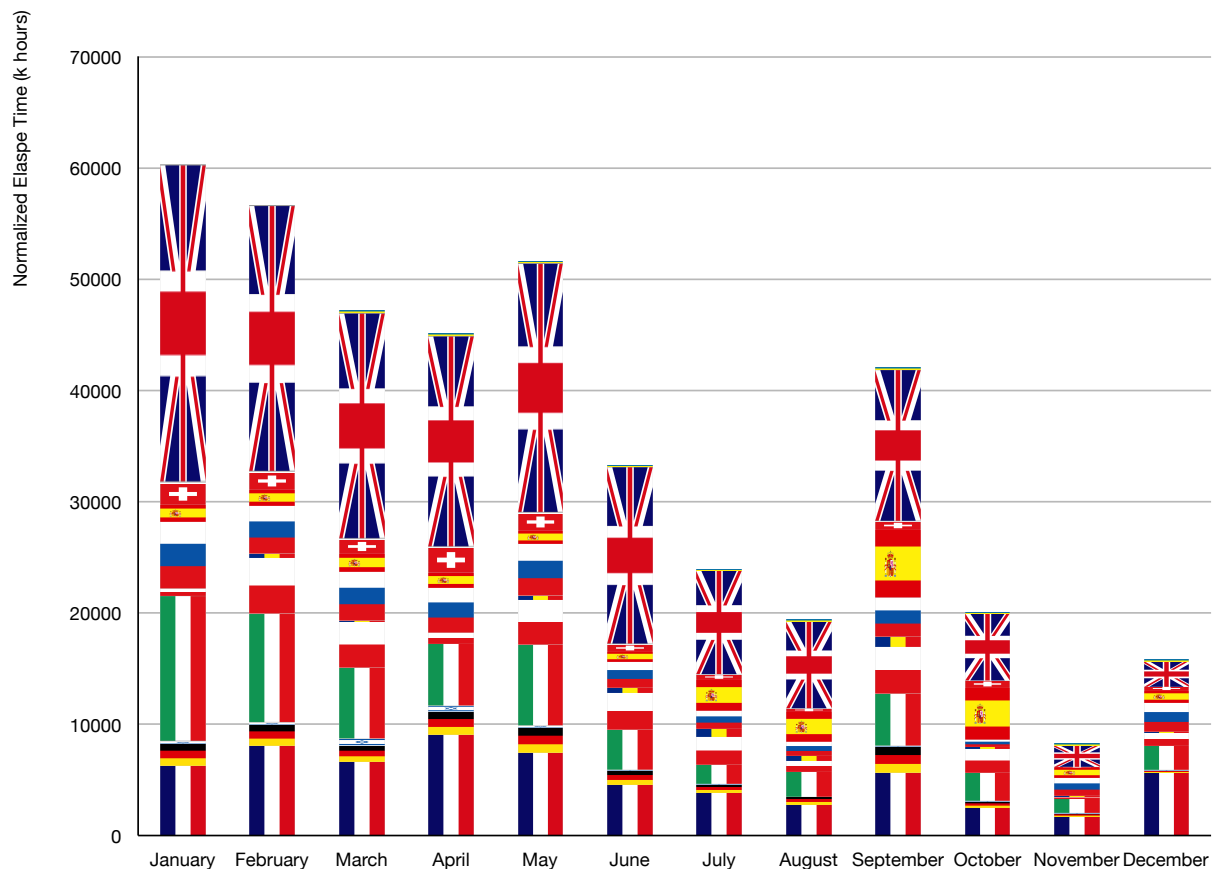




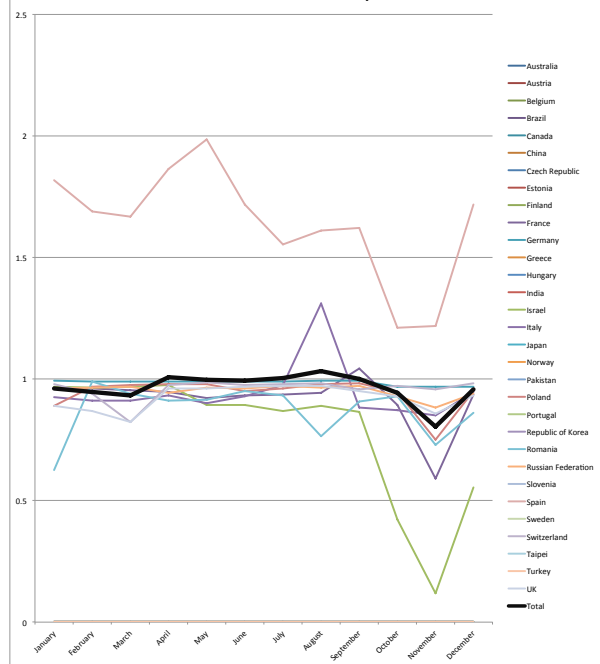
LHCb Plots 2011:



LHCB Tier-2 Normalized Elapse Time

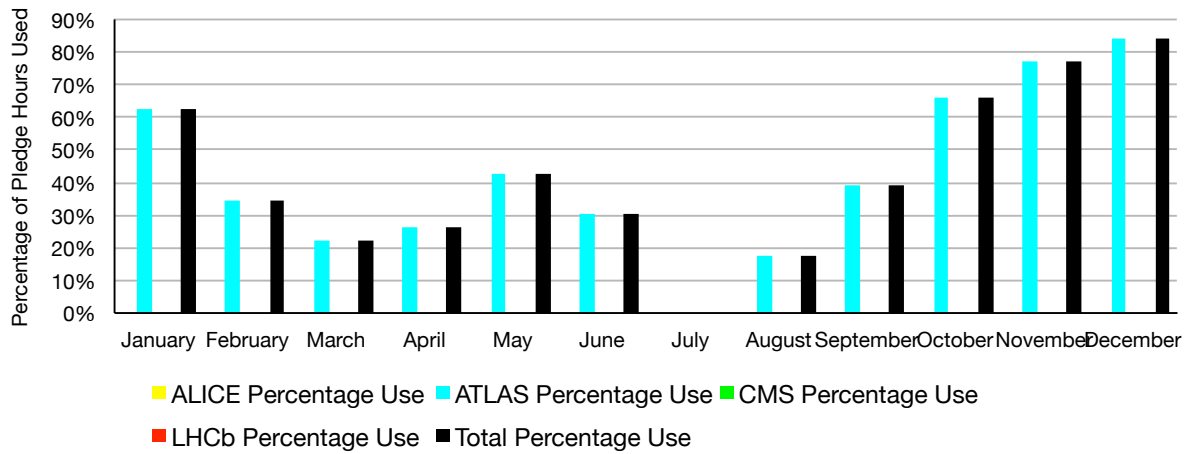


LHCB CPU Efficiency

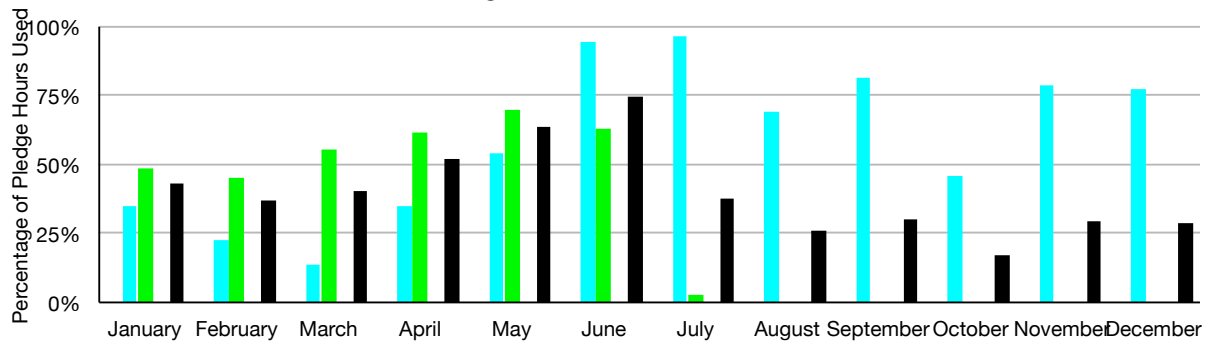


Tier-2 Country Plots 2011:

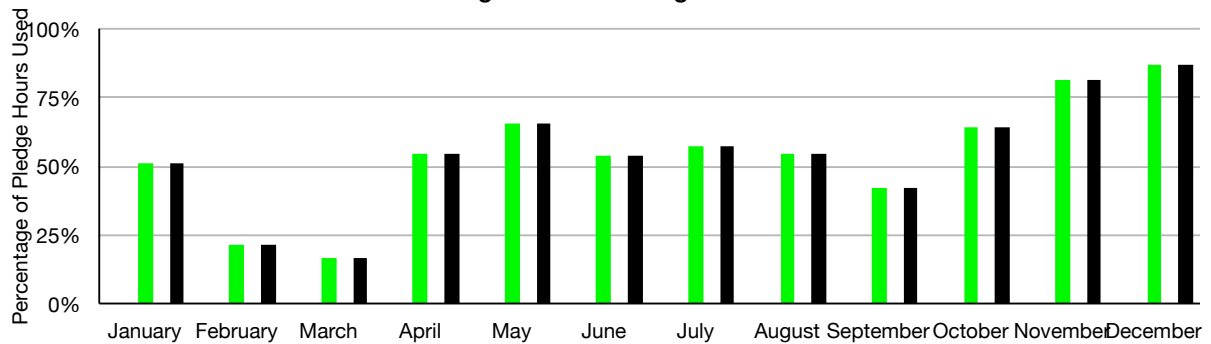
Percentage of Use for Australia Tier-2



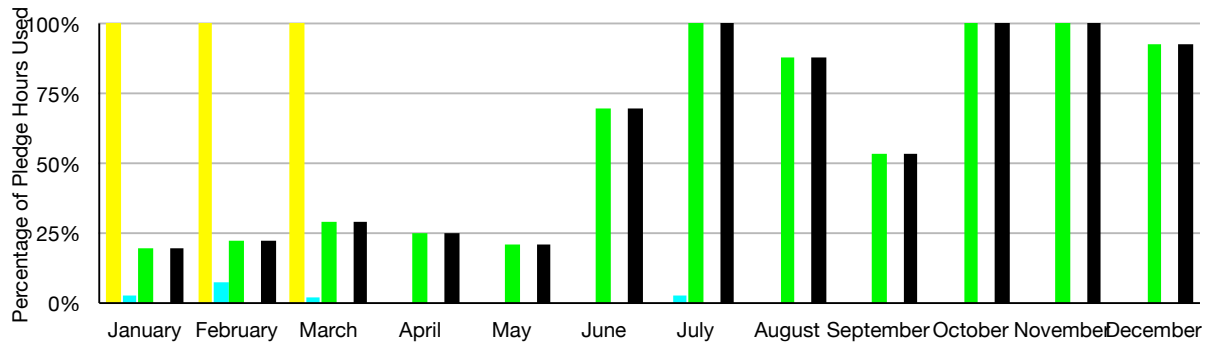
Percentage of Use for Austria Tier-2



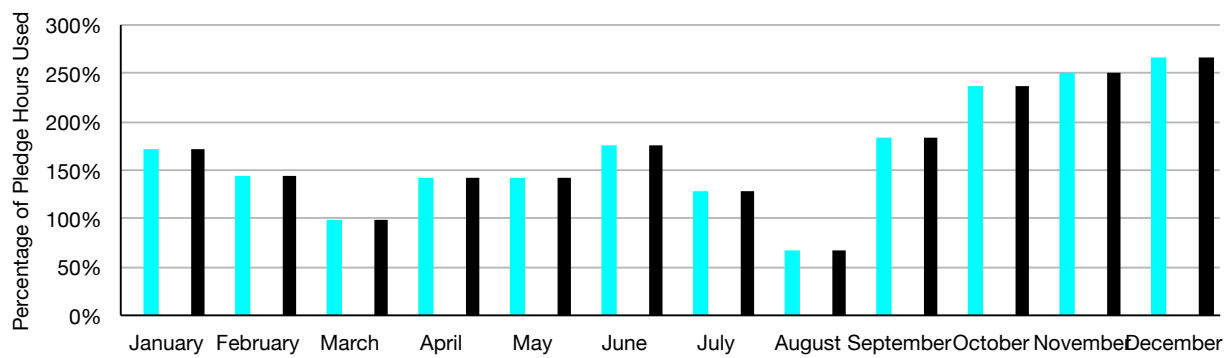
Percentage of Use for Belgian Tier-2s



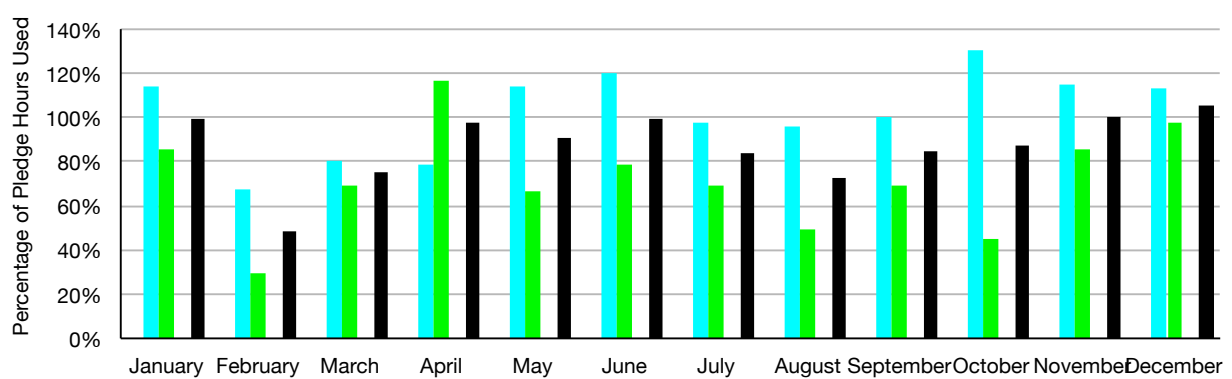
Percentage of Use for Brazilian Tier-2s



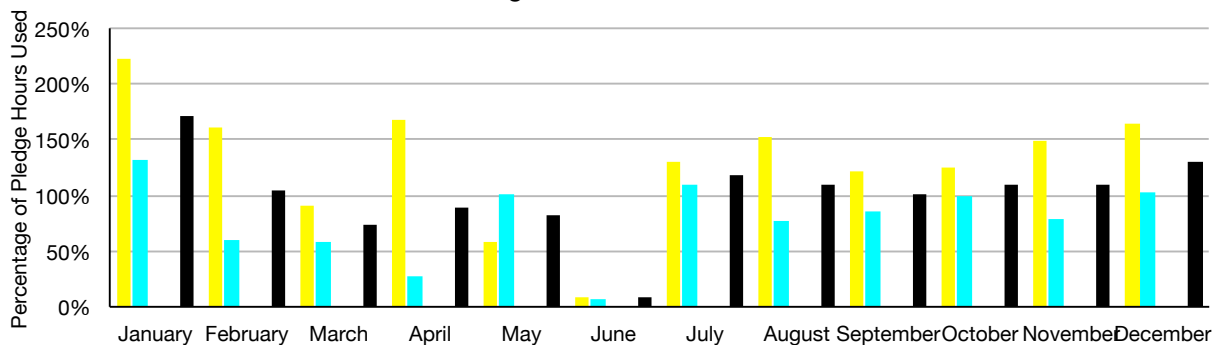
Percentage of Use of Canadian Tier-2s



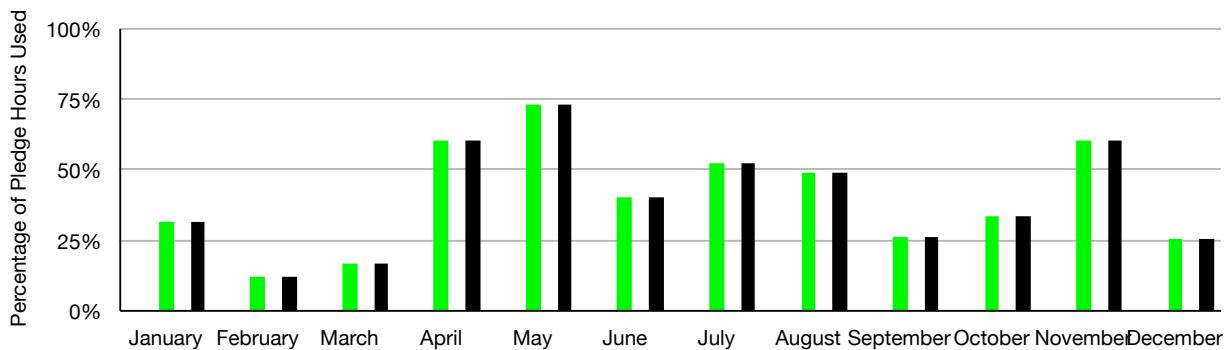
Percentage of Use of Chinese Tier-2s



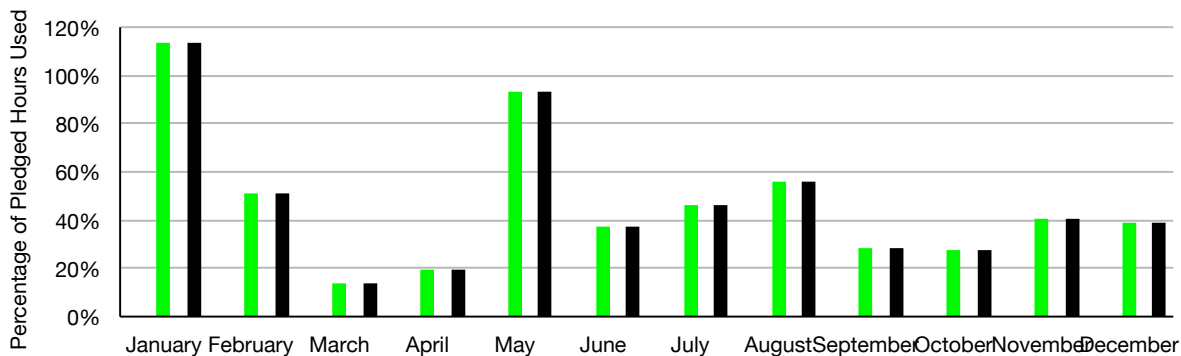
Percentage of Czech Tier-2s Used



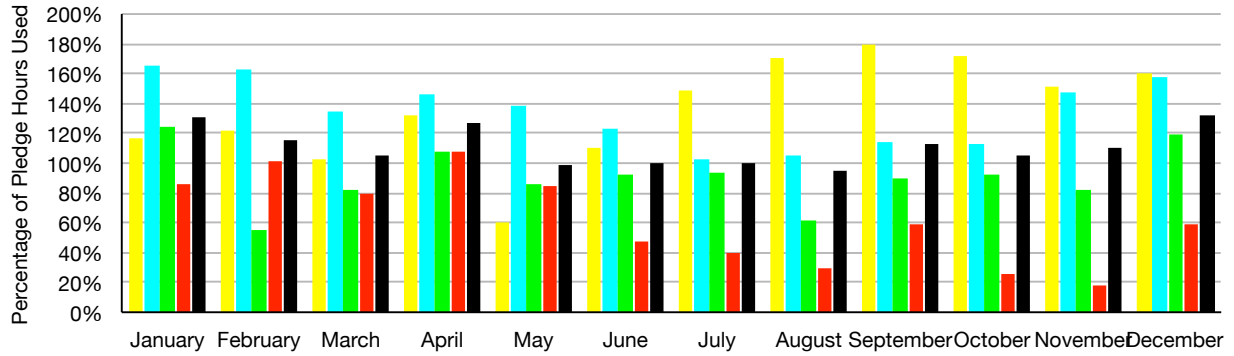
Percentage of Estonian Tier-2s Used



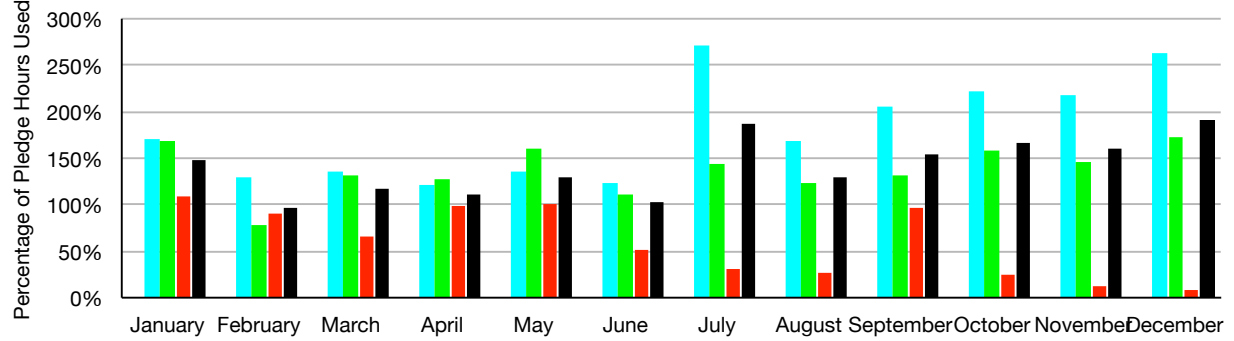
Percentage of Finish Tier-2s Used



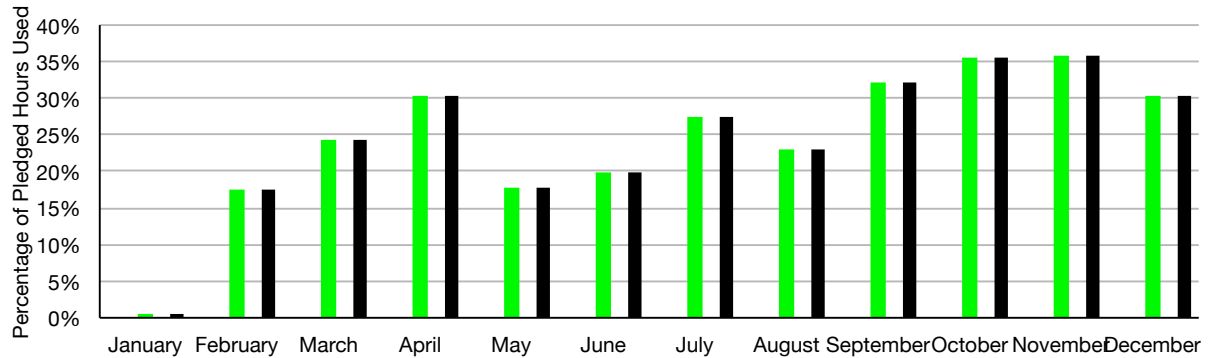
Percentage of French Tier-2s Used

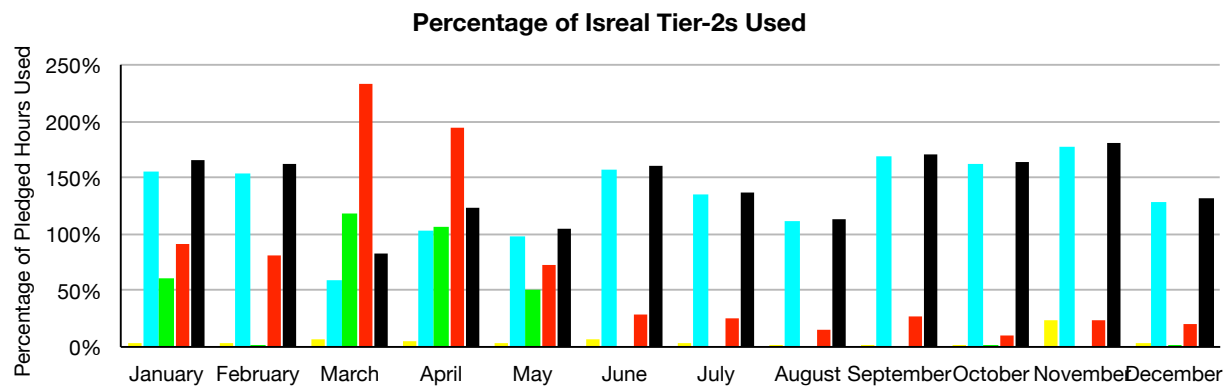
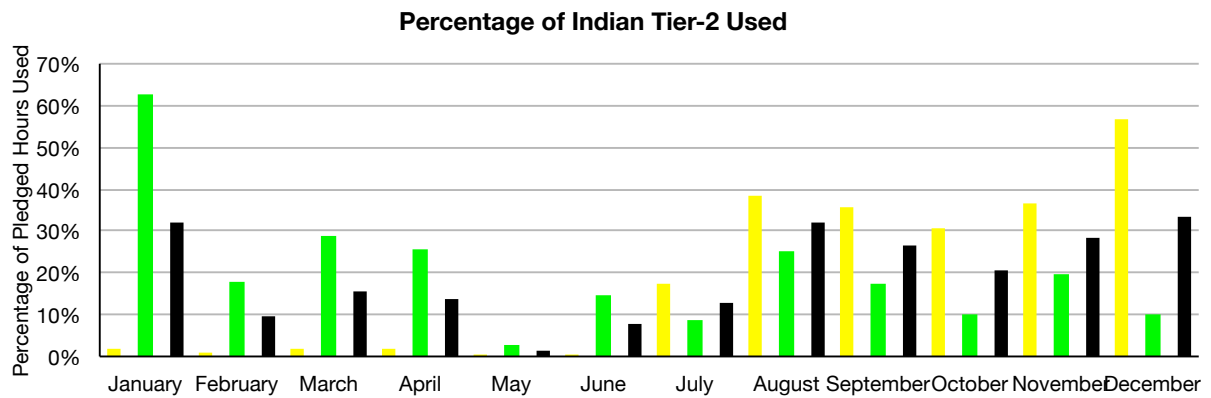
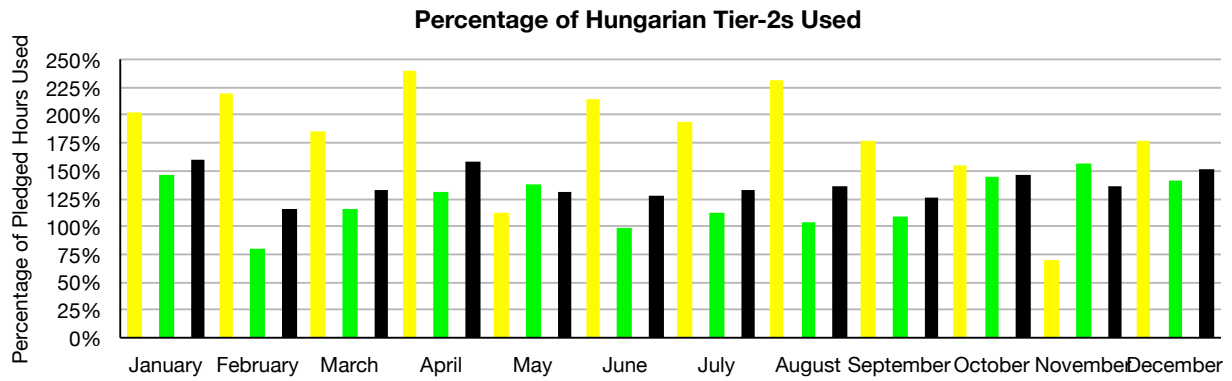


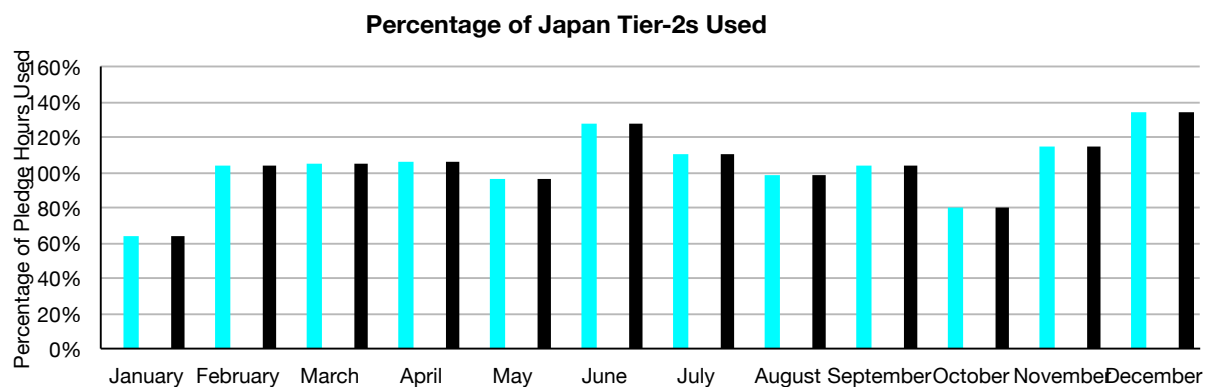
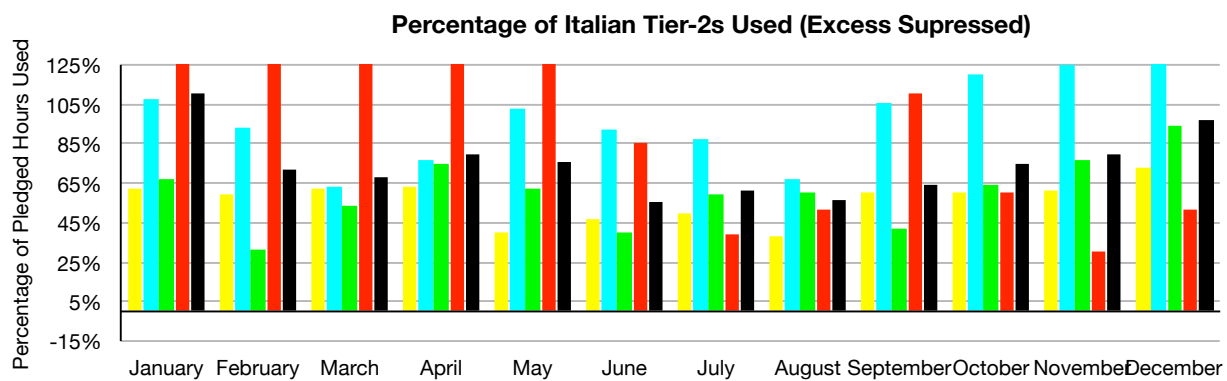
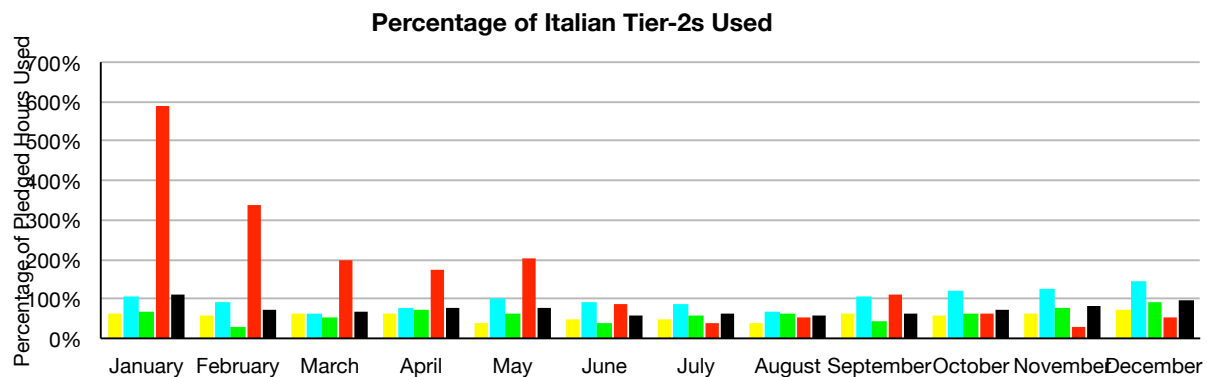
Percentage of German Tier-2s Used



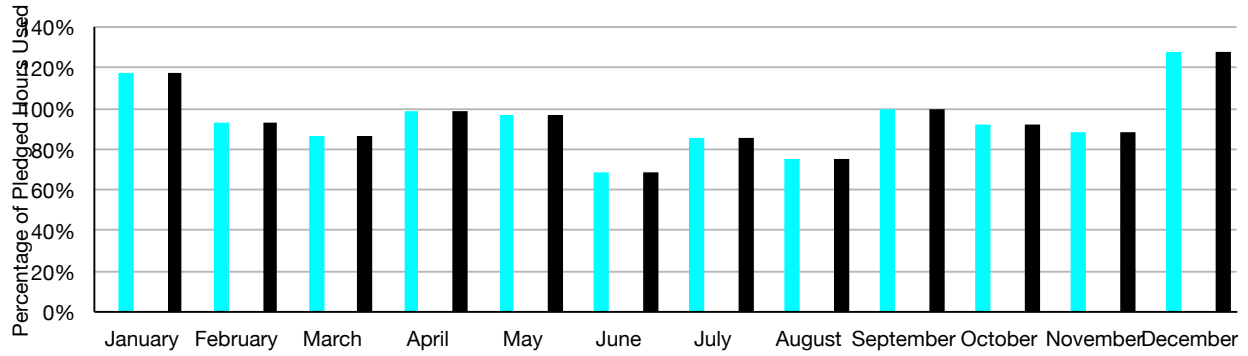
Percentage of Greek Tier-2s Used



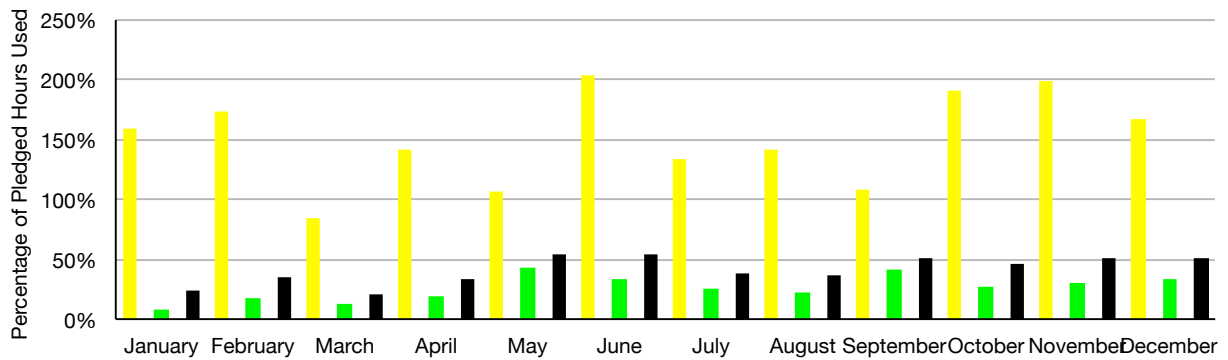




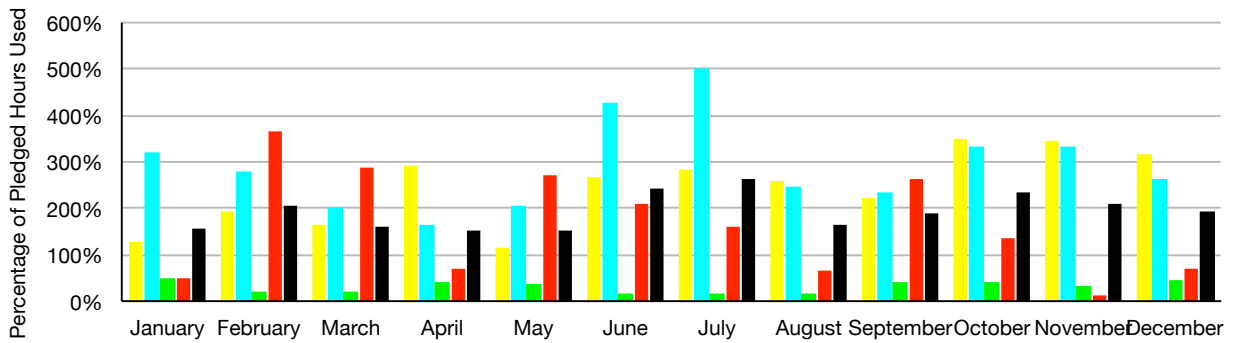
Percentage of Norway Tier-2s Used



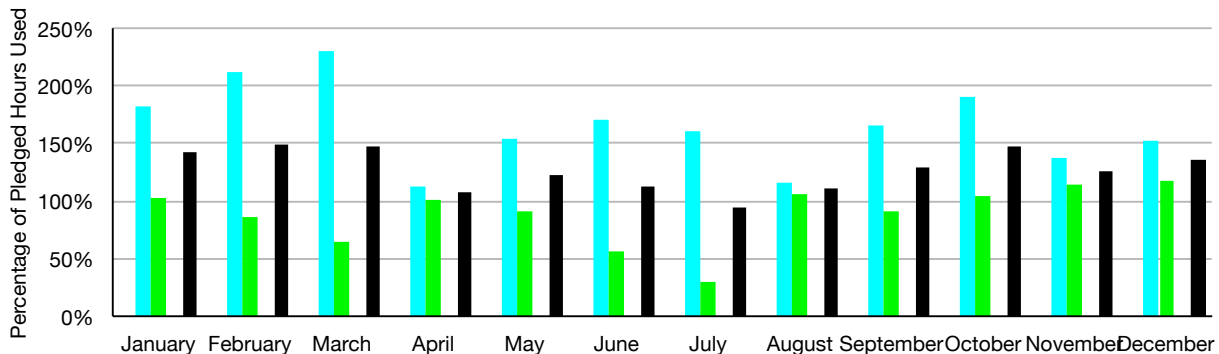
Percentage of Pakistan Tier-2s Used



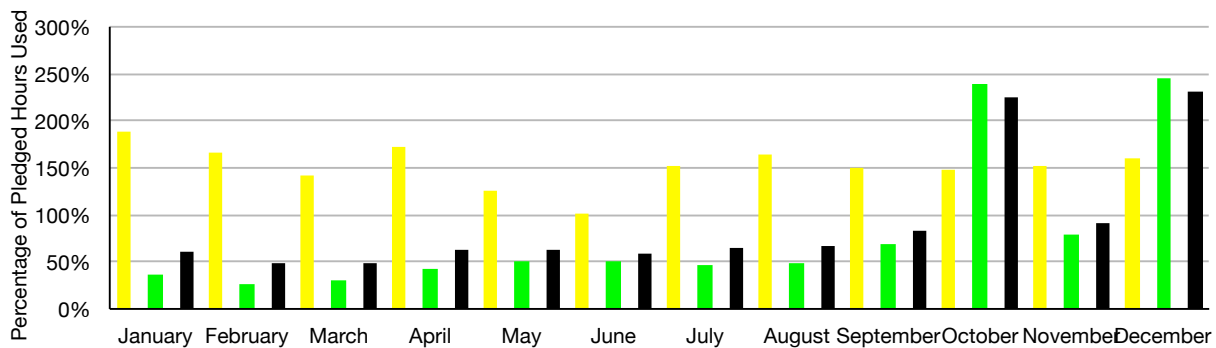
Percentage of Polish Tier-2s Used



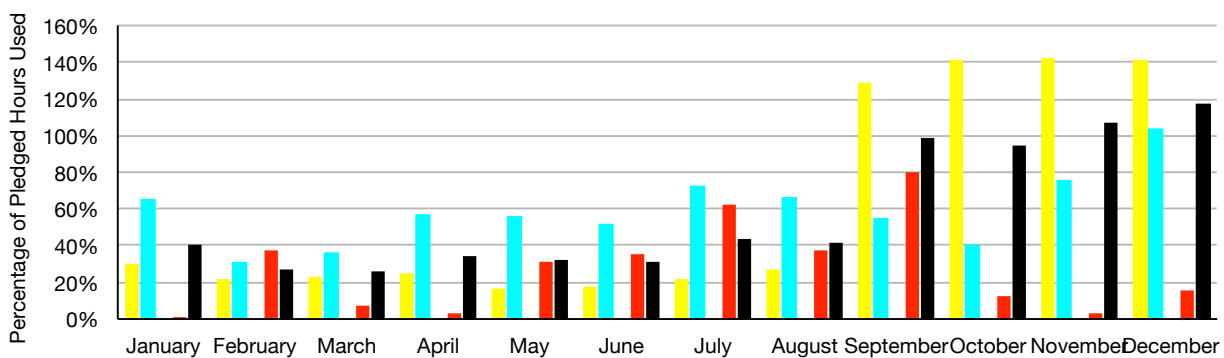
Percentage of Portuguese Tier-2 Used



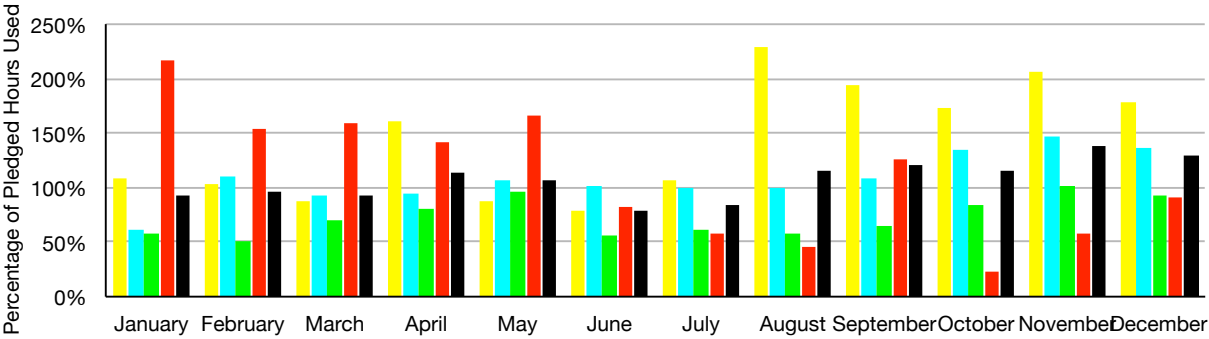
Percentage of Korean Tier-2 Used



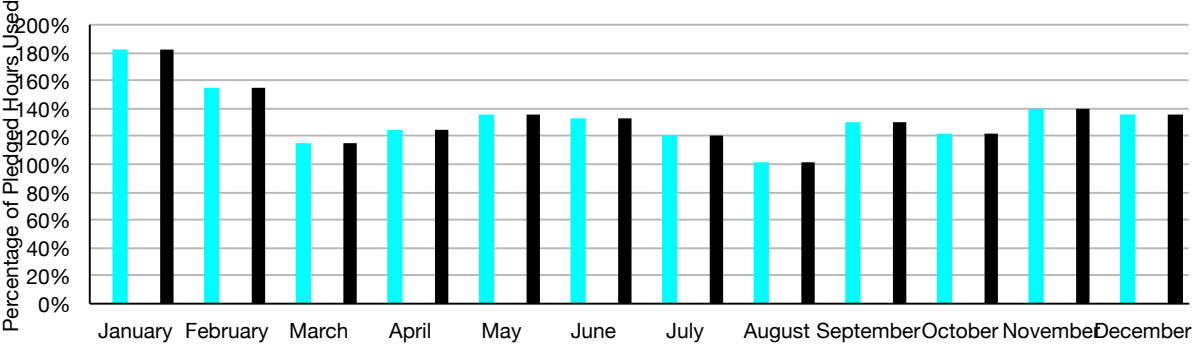
Percentage of Romanian Tier-2s Used



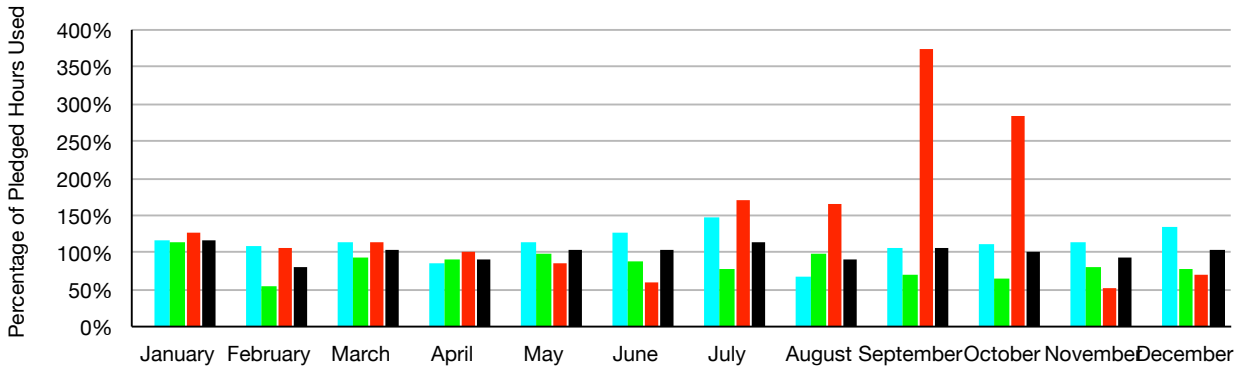
Percentage of Russian Tier-2s Used

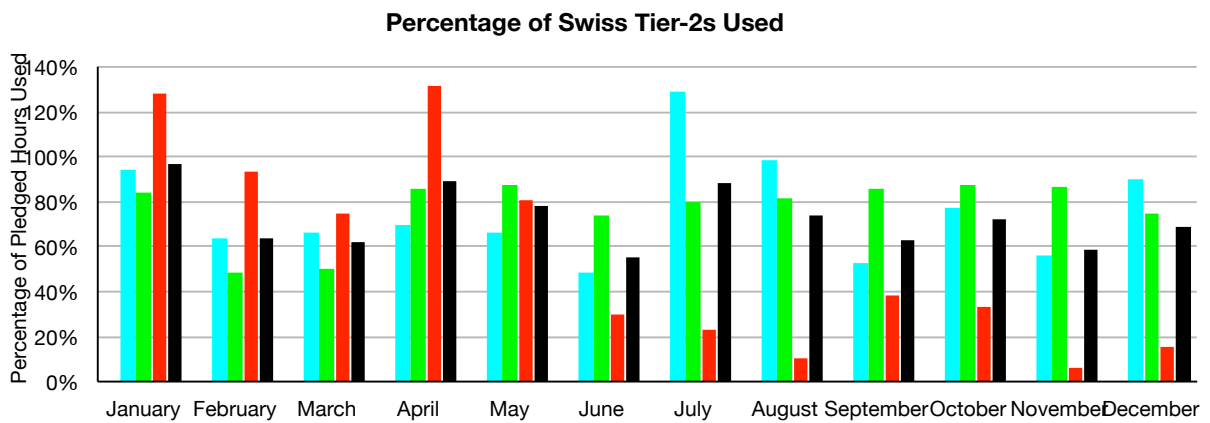
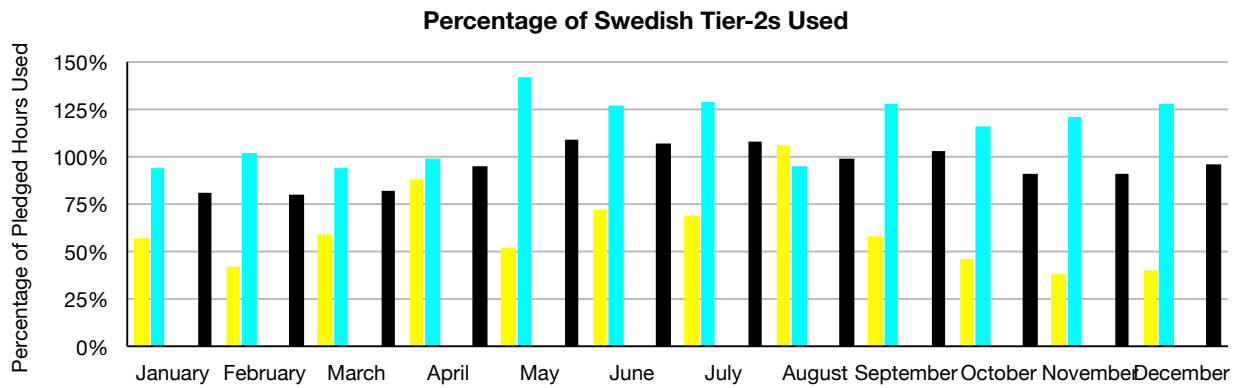
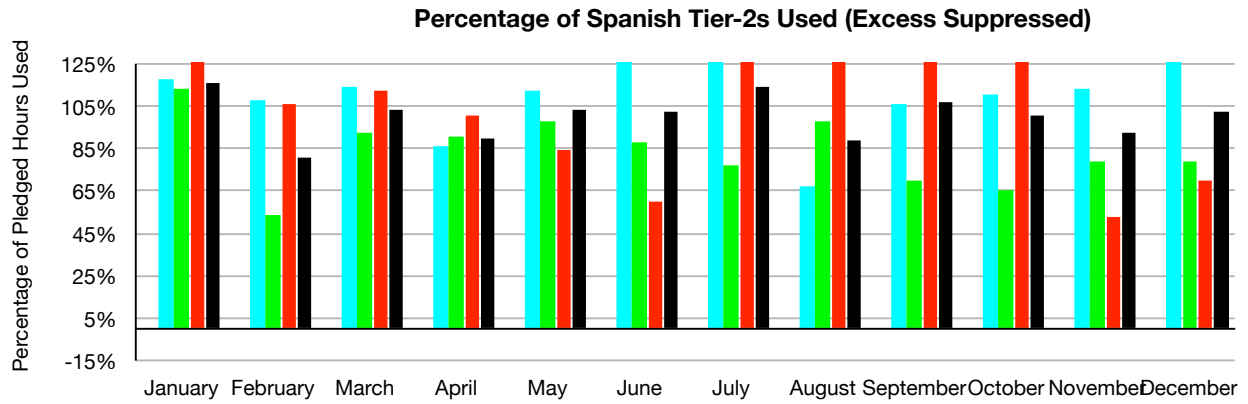


Percentage of Slovenian Tier-2s Used

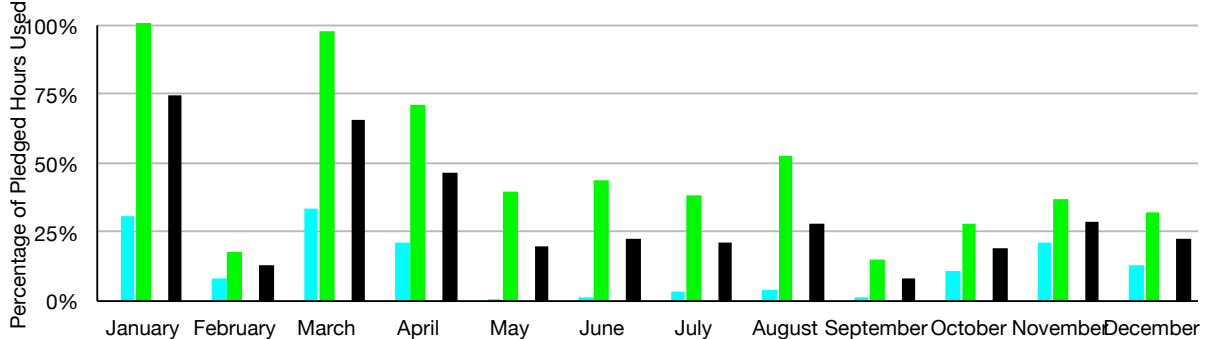


Percentage of Spanish Tier-2s Used

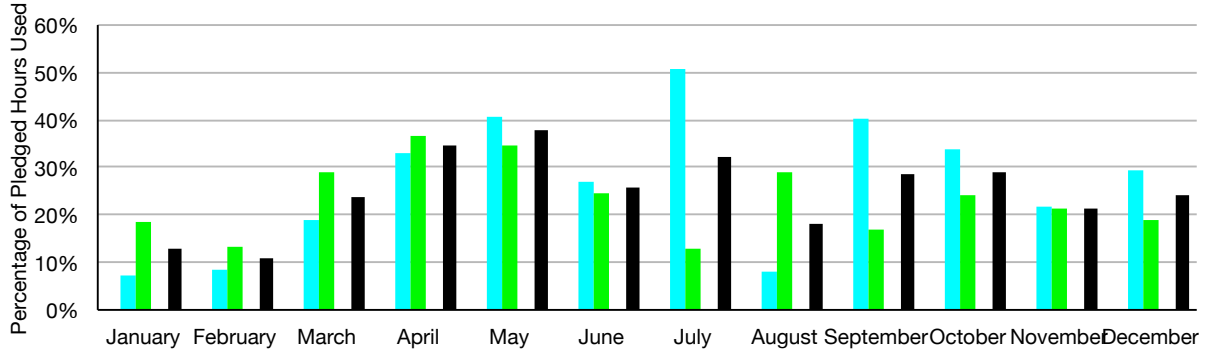




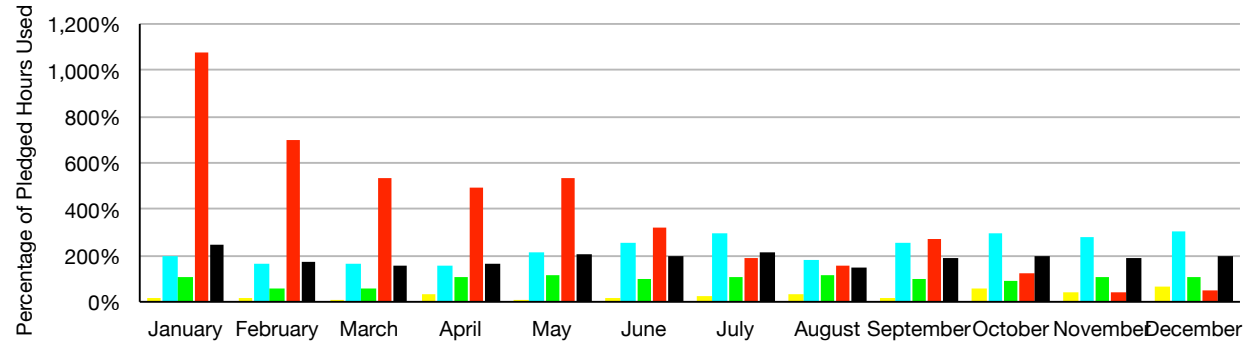
Percentage of Taipei Tier-2s Used

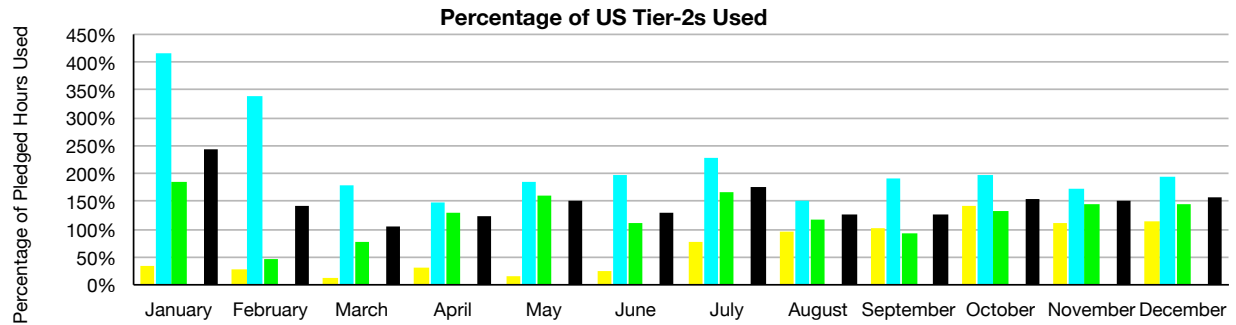
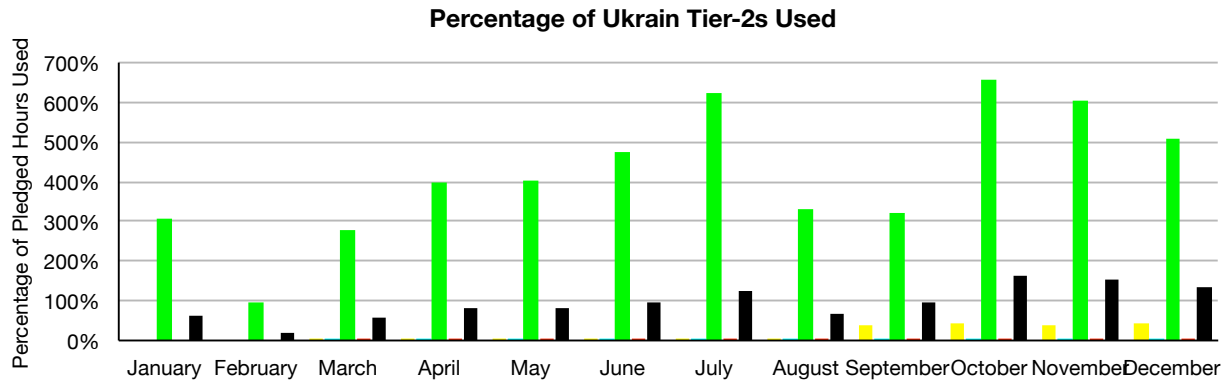
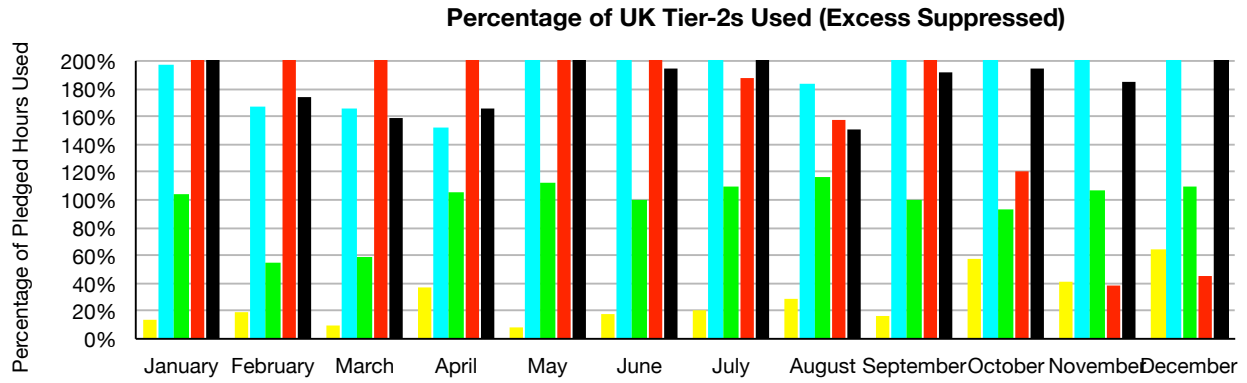


Percentage of Turkish Tier-2s Used



Percentage of UK Tier-2s Used





Total Tier-2s Used

