

# **C L R C**

RUTHERFORD APPLETON LABORATORY  
E-SCIENCE CENTRE

Version 1.0

*Tier1A Utilisation 2005*

M.P.Hodges  
R.A.Sansum

13 January 2006

---

## **1.INTRODUCTION**

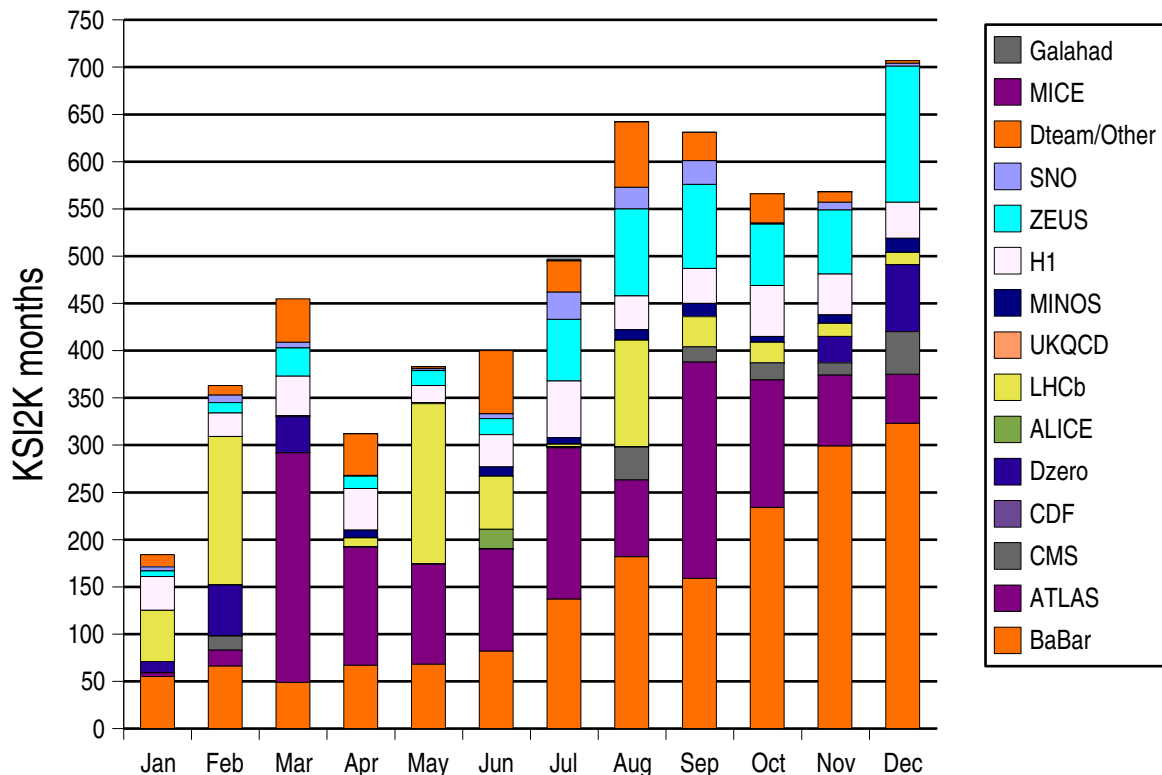
Low CPU utilisation during 1H05 was a concern to both the Tier-1 Board and Oversight Committee. This paper provides an update to the CPU utilisation data but also looks at occupancy (another measure of demand for resource) and job efficiency – which is believed to be part of the cause of lower utilisation than expected during Q2 2005.

## **2.CPU**

### **2.1CPU Utilisation**

By CPU utilisation of a batch job we mean the amount of CPU time (normalised between hardware generations to KiloSpecIntMonths) accounted to the job. This does not include system overheads such as I/O which are not accounted by the batch system. Nor does CPU time include time that a process spends in I/O wait waiting on network I/O. The nominal published farm capacity in July was 796KSI2K months per month. The available CPU capacity has been recalculated, and is currently 830KSI2K.

CPU utilisation has been essentially constant between February and June at just under 400KSI2K\*CPUMonths per month – about 50% of total capacity. Utilisation has been somewhat higher in the second half of the year, peaking at about 700KSI2K\*CPUMonths in December – about 85% of total capacity. The farm occupancies are larger than the percentage capacity utilisations, and full occupancy has been maintained for extended periods as discussed below.



**Figure 1: CPU Utilisation (KSI2K\*CPUMonths)**

Although superficially the periods of lower utilisation earlier in the year suggested that demand by the experiments for capacity has at times been low, the average job occupancy has been much higher, with the farm running at 100% capacity for periods in excess of 1 week. An analysis of the causes of this discrepancy is provided below.

## 2.2 Job Occupancy

By job occupancy we mean the total number of running jobs. Farm scheduling policy is to allow 1 job slot per CPU giving us at present approximately 890 job slots in total. In principle, all job slots can be filled simultaneously, however not all hardware is identical and demand from large memory jobs has effectively reduced total available slots as big jobs leave no memory available for second job starts. Since June 2005 the situation has been improved with additional queues for large memory jobs, however it remains the case that with the wrong job mix submitted not all job slots can be occupied.

Considerable progress has been made over the last 12 months. Farm occupancy (number of job slots occupied) has improved substantially since January. Following significant LCG load in August 2004 (off plot) much of the early month of activity were dominated by BaBar non-Grid activity until December 2004 when LHCb Grid work started running. Since January farm occupancy due to Grid based jobs has significantly increased, and these jobs are from an increasingly diverse range of VOs.

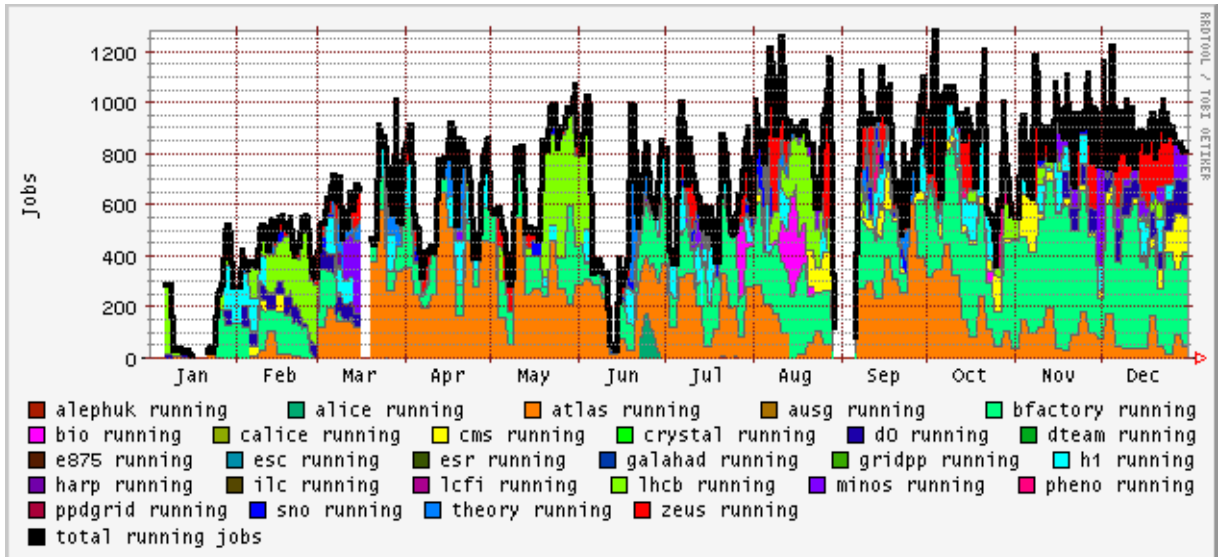


Figure 2: Farm Job Occupancy (2005)

Recently, farm occupancy has been at or close to 100%, and particularly encouraging has been the increasingly diverse job mix present. (Note tha occupancy above maximum capacity is believed to be an artefact of the way Ganglia gathers PBS job occupancy data.)

The occupancy for December 2005 is shown in the plot below :

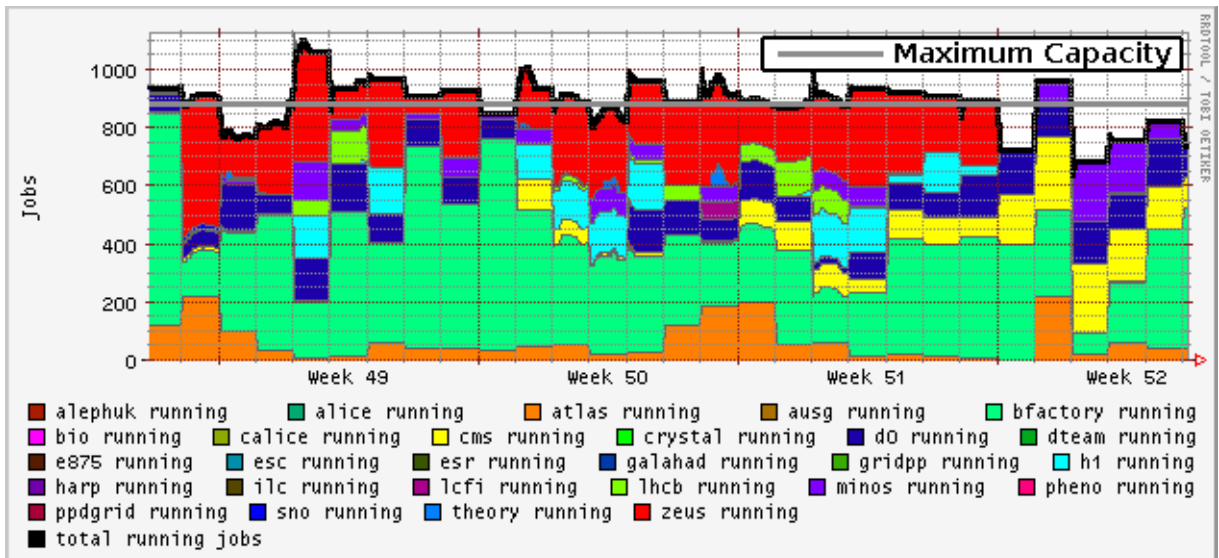


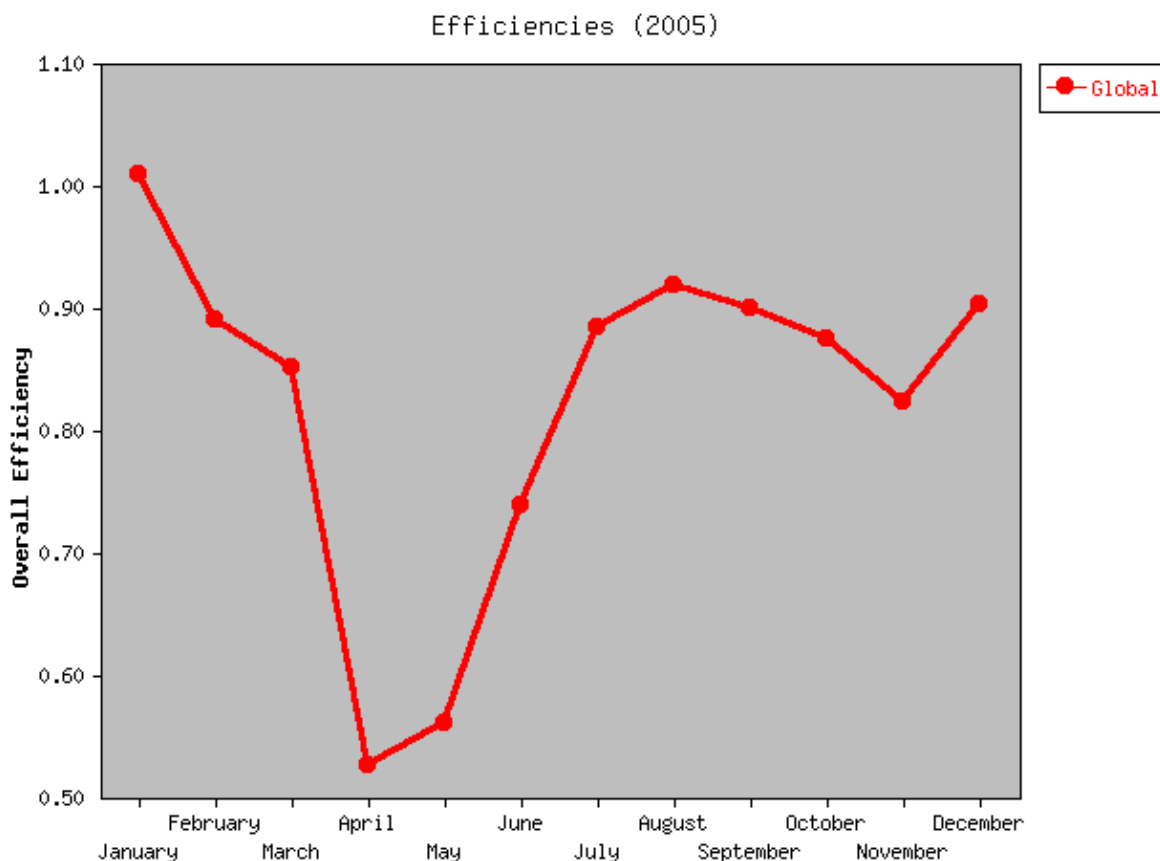
Figure 3: Farm Job Occupancy (December 2005)

This includes the Christmas period, where the farm was running “at risk”, and the level of maintenance was minimal. Despite this, occupancy remained high, and a full service was maintained.

## 2.3 Efficiency

The efficiency of a batch job is the ratio of CPU time to elapsed time. A highly CPU intensive batch job can achieve 95-98% utilisation of a CPU, an I/O intensive job is more likely to be around 85-95% utilisation of a CPU, and jobs stuck on busy resources can vary from 0-100%. It is possible for a job to have an efficiency greater than 1.0 if it runs more than 1 CPU intensive process - few jobs are in this category. The global efficiency of the farm is defined as the ratio of the overall elapsed CPU time to wall clock time for all experiments.

A detailed analysis of batch job efficiency has been carried out. As can be seen overall efficiency was very low during Q2/2005, but was much higher in Q3/2005 and Q4/2005, and was 90% in December.

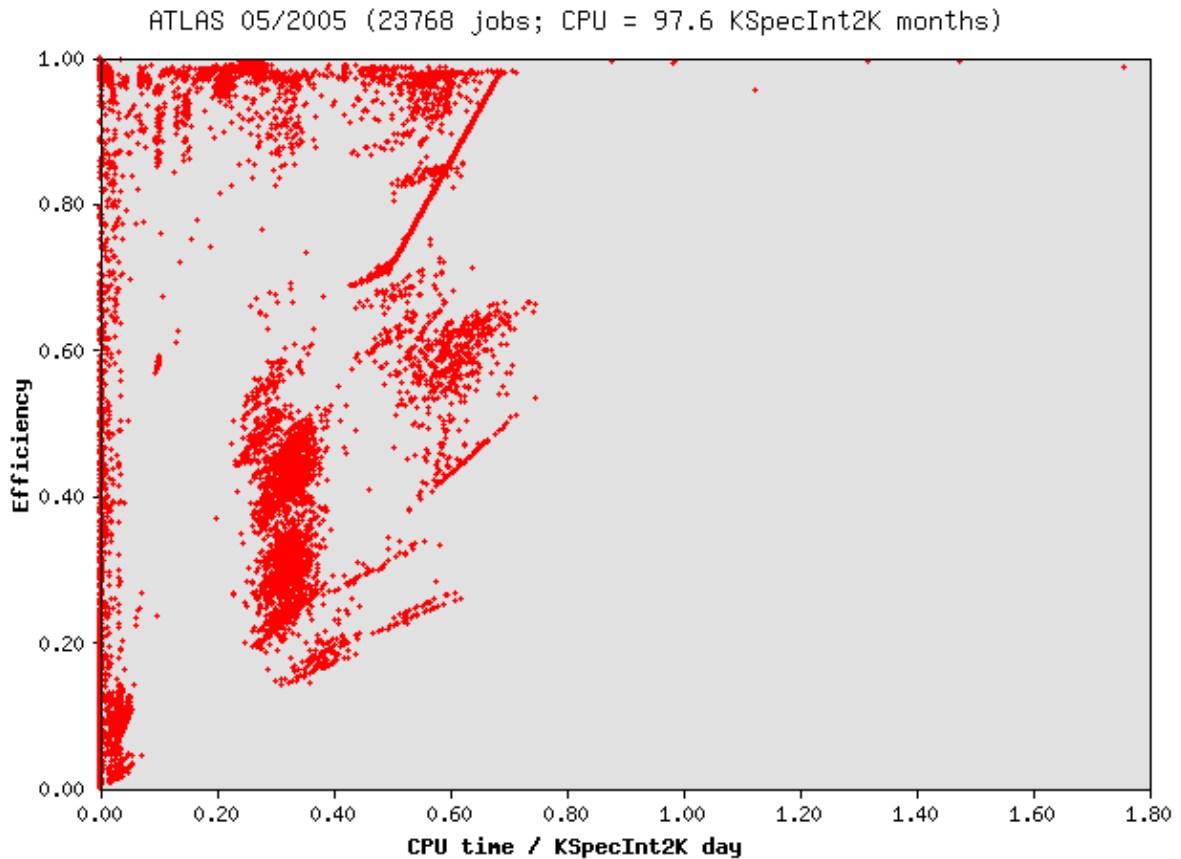


**Figure 4: Overall Job Efficiency (2005)**

Most of the efficiency problems were related to global performance problems on LCG storage elements, coupled to problems in logging and book-keeping services. Local I/O bottlenecks were also thought to be a contributing factor to low efficiencies.

An example where efficiency has been a problem is shown below where the scatter plot shows efficiency of individual jobs. As can be seen, a number of straight line structures show jobs which ran for a period of time before blocking on an external resource before eventually being

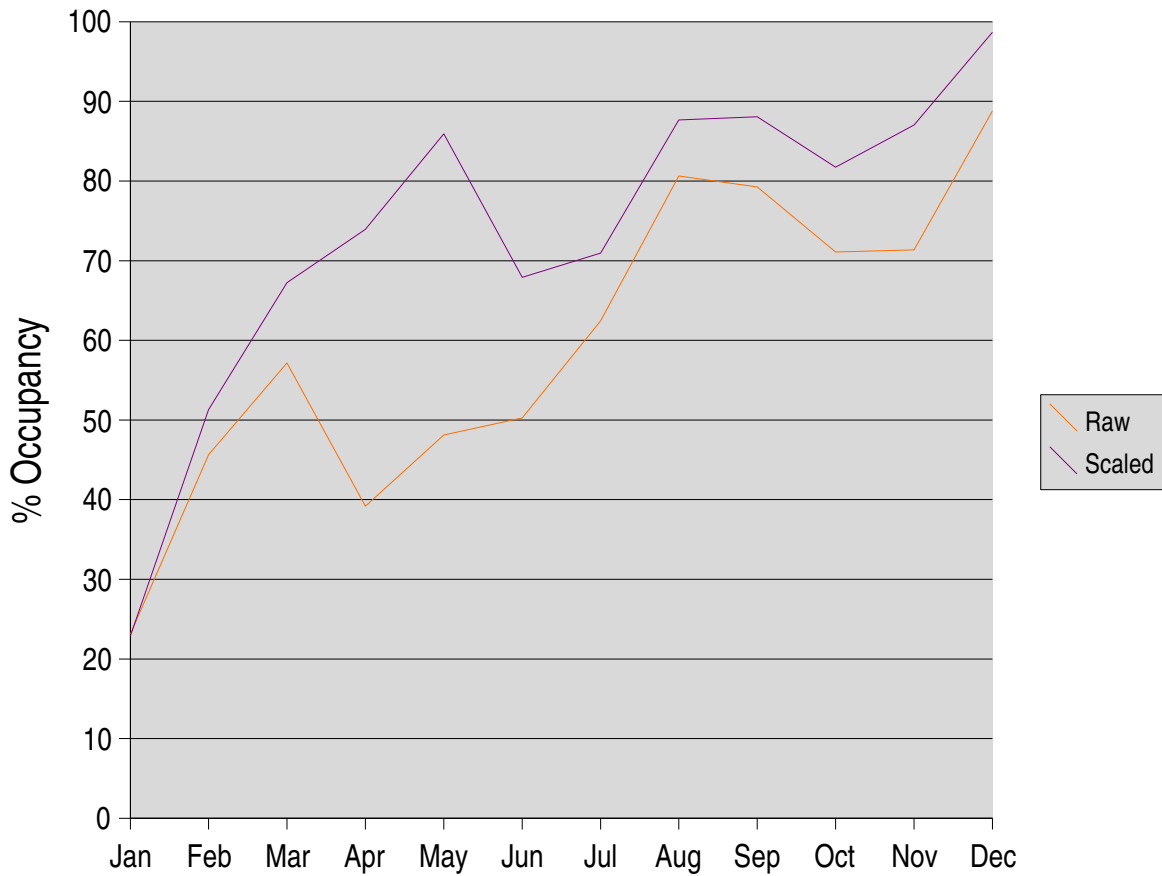
killed by an elapsed time limit. Other clusters at low efficiencies probably indicate performance problems on external storage elements.



**Figure 5: Individual Job Efficiency (ATLAS experiment – May 2005)**

A number of efficiency related problems have now been addressed by the LCG experiments and efficiency has been 90% or higher in August, September and December. We expect further improvements will be made on some of the existing efficiency problems; however as capacity and thus the scale of computing continues to climb, it is likely that further efficiency problems will be encountered. This remains part of the reason for the continued requirement to ramp up peak capacity.

When the utilisation data is corrected by the efficiency we get an estimate of what farm occupancy would have been delivered if the raw data were scaled for 100% efficient jobs. This is shown below:

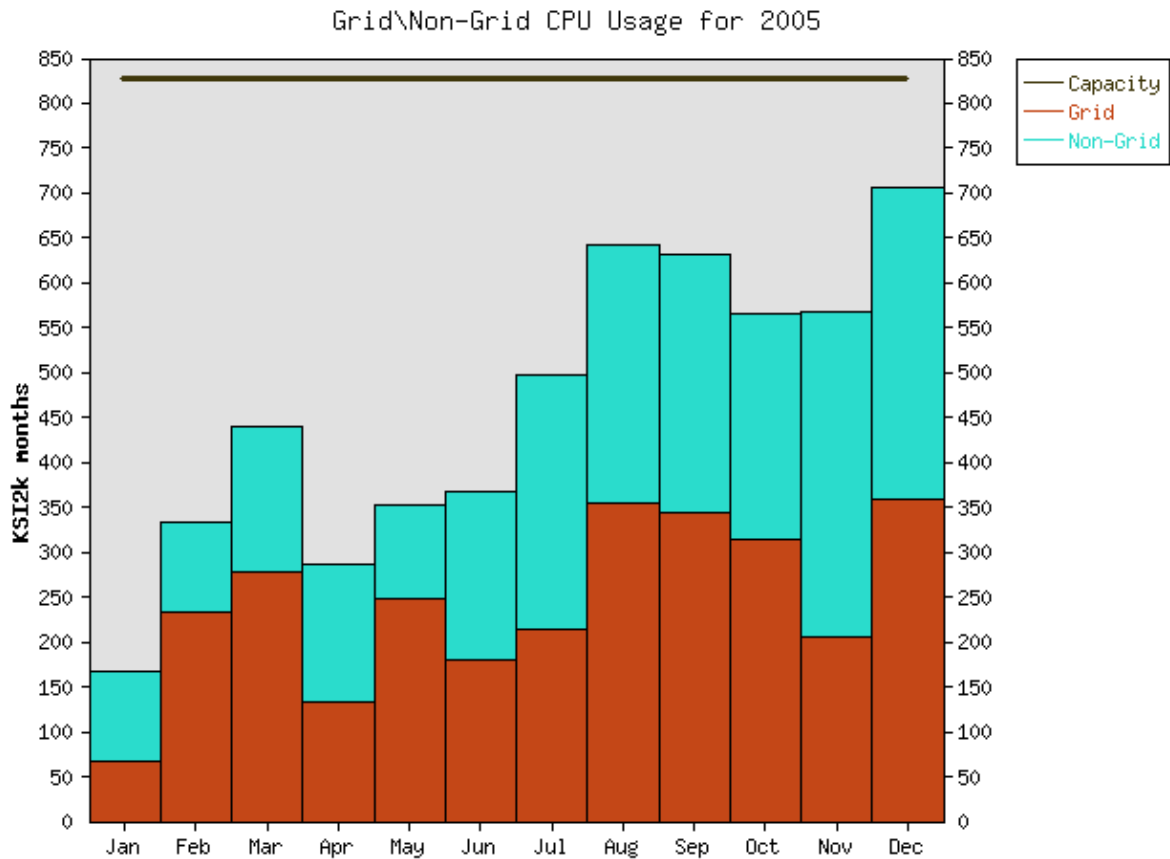


**Figure 6: Estimated % Occupancy**

Occupancy rose dramatically during the early part of the year, the scaled value has fluctuated around 80% after the first few months and reached near to 100% occupancy in December.

## 2.4 Grid v.s Non-Grid Usage

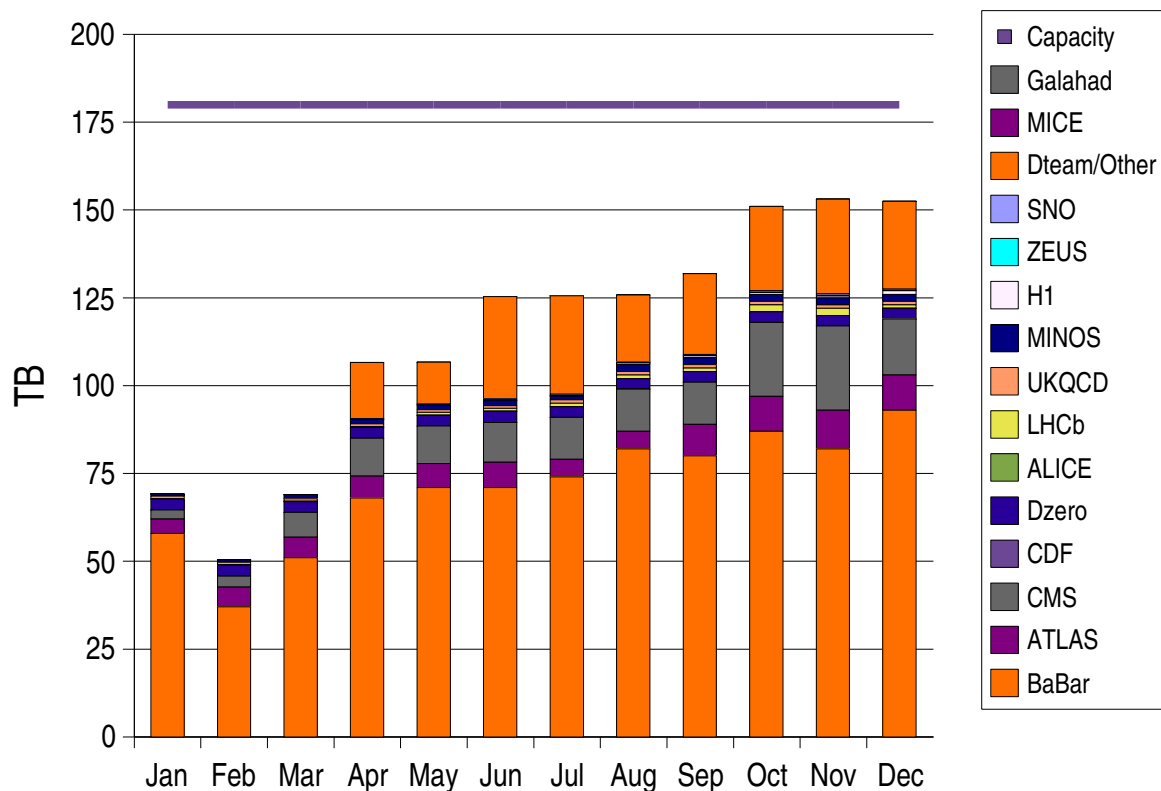
An analysis of CPU usage shows a general increase of Grid usage over 2005. Non-Grid CPU has also increased, and the proportion of Grid work has fluctuated around the 50% mark over this period.



As long as BaBar analysis work is not run via the Grid, the percentage of Grid jobs is not likely to be significantly higher than seen in 2005, provided that BaBar run at allocation, which for Q1/2006 is about 50% of farm capacity.

### 3.DISK CAPACITY

Current available disk capacity is about 180TB. Disk utilisation has climbed steadily over 2005 and is currently about 150TB (including Service Challenge capacity), BaBar is the dominant user of disk space and they usually have an excellent record of using what they have been allocated.



**Figure 7: Used Disk Capacity (TB)**

Unused capacity available to end users was therefore about 30TB in Q4/2005.

Although it is clear that there remains some unused capacity it should be noted that headroom is usually needed by the experiments during normal operation. For example, BaBar retains about 10% capacity unused in order to operate a dynamic pool for dynamic restore of offline datasets stored on tape. File-systems with more fluid storage policies need rather lower occupancies in order to avoid failures due to lack of space. This optimal occupancy will vary between experiments but experience suggests that 70-90% is not unreasonable, depending on the data rates and size of the disk pool (for the service challenge we aimed for 20% headroom to ensure acceptable filesystem performance and to provide adequate buffer space against rate fluctuations). Current overall utilisation is about 80% and there probably remains headroom of 20-30TB.

The main obstacles to full utilisation have been:

- Problems achieving excellent reliability of the Storage Element (SRM)

- Late deployment in LCG of the File Transfer Service.

There remain some problems in both these areas, however substantial progress was made during service challenge 3 over July.

#### 4.TAPE CAPACITY

Tape capacity has remained constant for most of 2005, at 340TB. BaBar remains the largest user of tape.

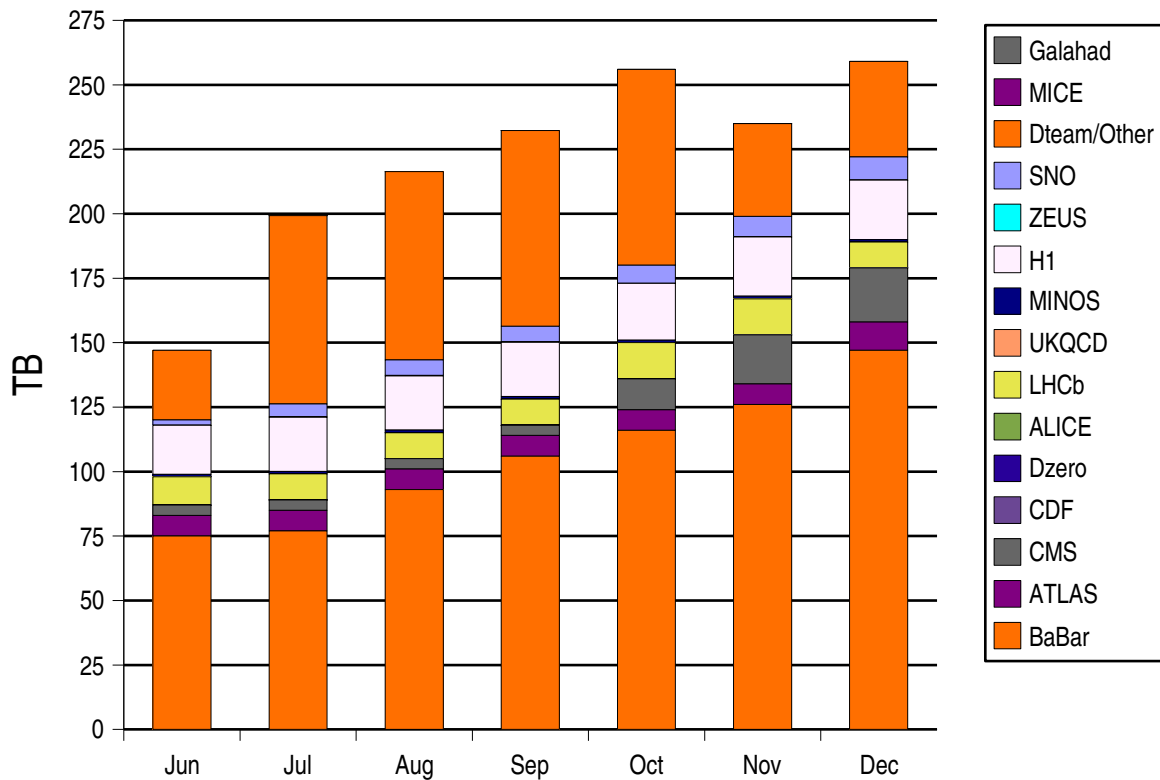


Figure 8: Used Tape Capacity (TB)

#### 5.CONCLUSIONS

Over 2005 the Tier-1 has run an increasingly vibrant mix of experiment jobs as more and more VOs have become active on the Grid. The number of jobs submitted by the Grid has also increased substantially in this period. Job occupancy has increased dramatically since the start of the year and, although there were quieter periods where demand (globally) was low, there are increasingly long periods where the service is running at 100% occupancy. Key players such as ATLAS and LHCb have sustained Grid production work over most of 2005 (RAL was a major LCG Tier-1 contributor to both ATLAS and LHCb production), BaBar work was impacted by the SLAC shutdown but is now running close to allocation.

Low Grid job efficiencies, especially in Q2/2005, have at times considerably reduced CPU utilisation and prevented the Tier-1 delivering near to its maximum theoretical capacity. A number of these issues were resolved, and during the first part of August the farm sustained several weeks at almost full CPU capacity for the first time. Since August, sustained periods of running at full capacity have become more common. Problems in the disk SRM and LCG File Transfer Service have deterred take up of disk capacity by LCG experiments, but LCG considers this a priority area and work continues to resolve these issues.