Optimised access to user analysis data using the gLite DPM

Sam Skipsey s.skipsey@physics.gla.ac.uk
Graeme Stewart g.stewart@physics.gla.ac.uk
Greig Cowan gcowan@ed.ac.uk

ScotGrid Scottish Grid Service

GridPP UK Computing for Particle Physics

Introduction

There are two differing use cases for clusters supporting the WLCG VOs. Production use is well understood, as production jobs have run on the Grid for several years, and the existing infrastructure and best practise has evolved to optimise performance for them. User analysis is, by comparison, poorly understood: there is no 'real' user analysis possible until the LHC actually starts producing data, and test simulations of user analysis access patterns have only recently begun against Tier-2 sites. It is widely accepted that the problem of providing efficient resource for generic user jobs is not solved, and that their access patterns are significantly different to those of the existing production jobs.

ATLAS Analysis + HammerCloud

The HammerCloud framework has recently been developed to allow automated submission of large numbers of user analysis jobs to ATLAS Tier-2 sites using the ganga distributed analysis tool. This allows Tier-2 sites to test their performance against load from a sample ATLAS user analysis, in a repeatable manner, while also providing metrics of success for comparison. Along with several other UKI Tier-2 sites, UKI-SCOTGRID-GLASGOW has used regular HammerCloud analysis tests, consisting of approximately 300 jobs per site, to test the performance of their storage infrastructure.

Initial testing (top-right quadrant) showed that performance of the DPM storage was noticably lacking, barely achieving event-rates of 10Hz, with a concomitant limitation on the mean and maximum job efficiency. The load on the DPM head node itself implies that it is clearly the effective bottleneck in this case.

Optimisation Steps

Two main optimisation steps were undertaken. In the first step (bottom-left quadrant), the DPM services were separated from the backend MySQL database server, and migrated to a newer host with significantly greater performance (from dual single-core CPUs to dual quad-core). In the second step (bottom-right quadrant), the MySQL databases were optimised by the application of several indices on common lookups (the cause of the 'noisy' CPU load plot in the previous data) and the enlargement of the bufferpool space for the database engine. Full details of optimisation work are available at:

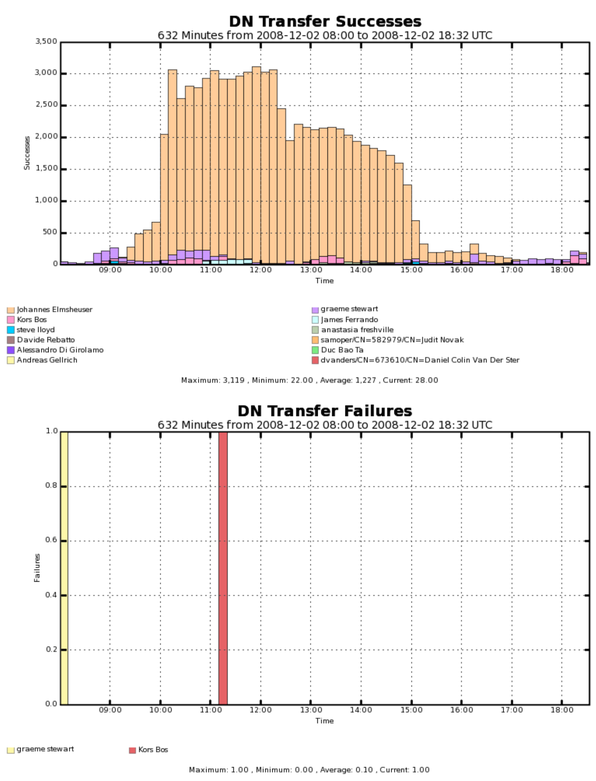
http://www.scotgrid.ac.uk/wiki/index.php?title=DPM_Optimisation

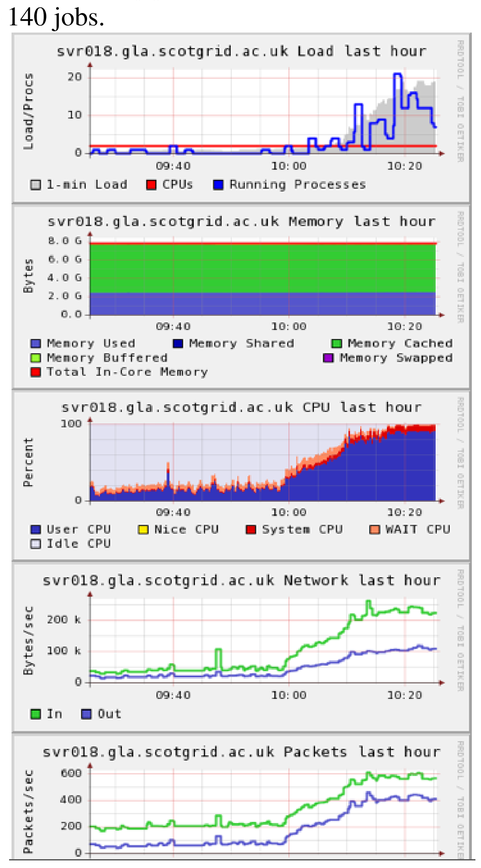
Acknowledgements

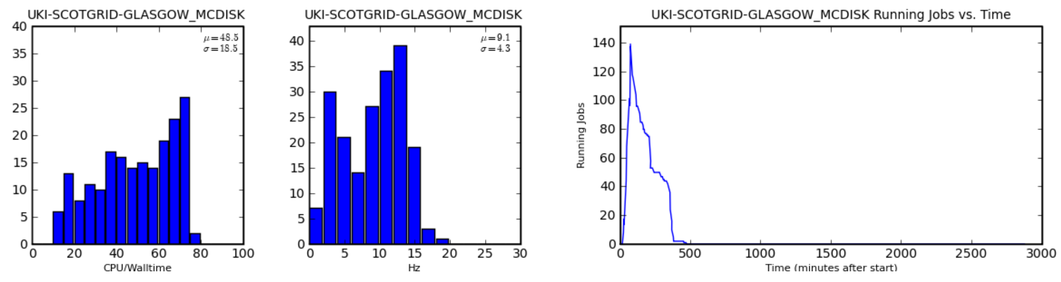
Johannes Elmsheuser and Dan Vanderster for HammerCloud testing framework and test scheduling.

HammerCloud Test 038 (before any optimisations)

Very high load on DPM head node; resultant event rate < 9Hz, many job failures due to timeouts. Peak concurrent load - 140 jobs.

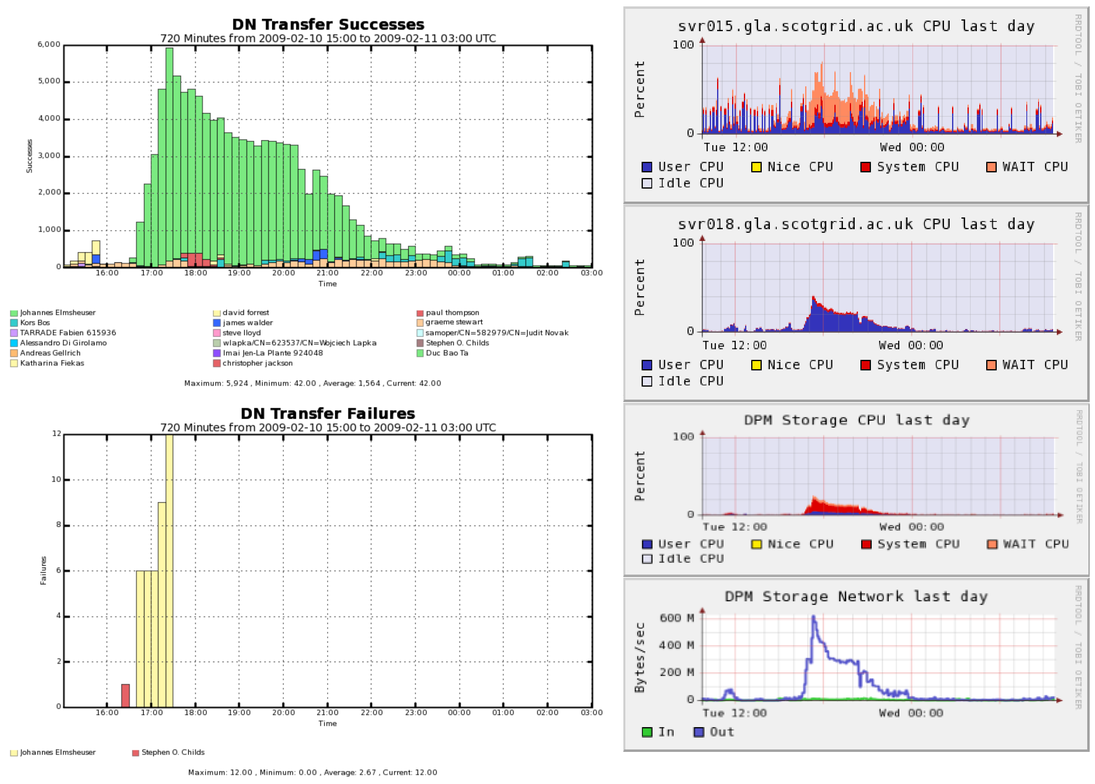






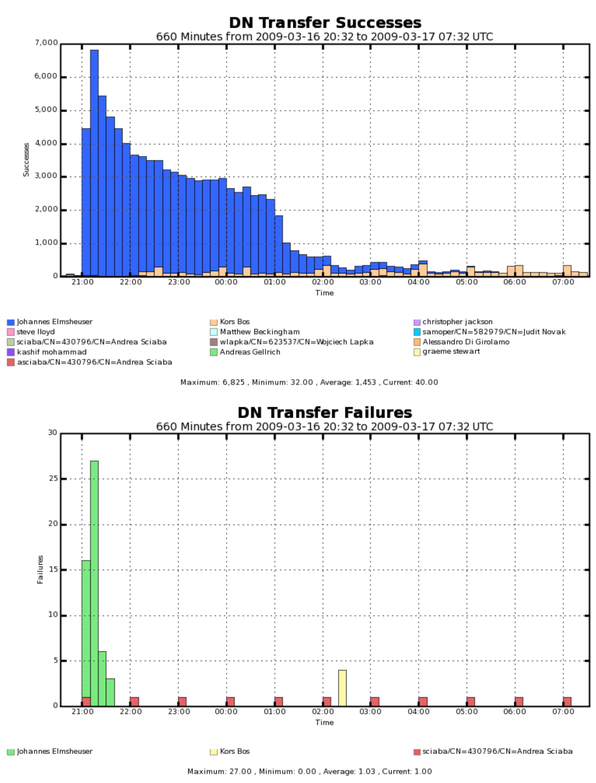
HammerCloud Test 135 (after splitting MySQL and DPM)

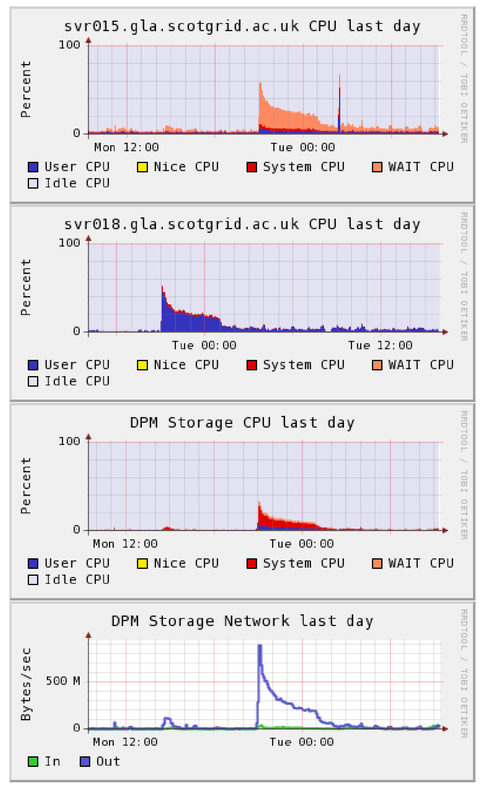
Much improved load on DPM and MySQL head nodes, with mostly IOwait on the MySQL node. 50% increase in eventrate and concomitant increase in efficiency. Peak concurrent load - 200 jobs.

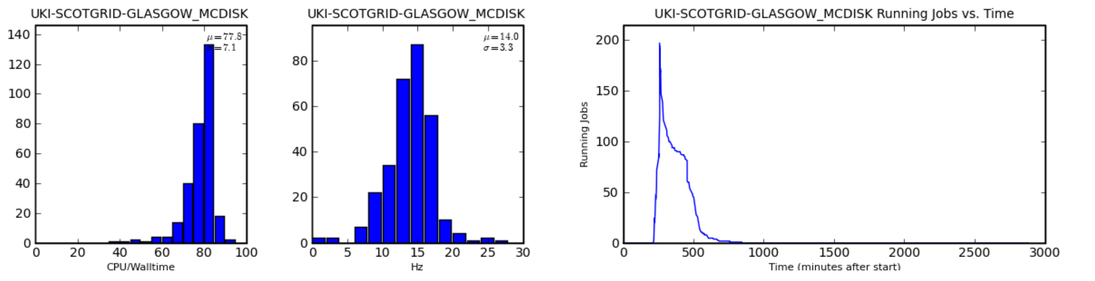


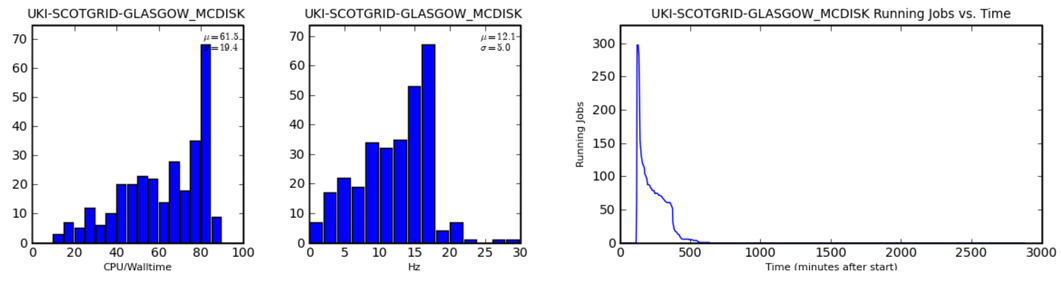
HammerCloud Test 193 (after all optimisations)

Further reduced load on MySQL node; all background load noise removed. Event rate slightly reduced due to the increased number of concurrent jobs in this test - 300 peak, compared to 200 in the previous test. Rate limiting may increase overall efficiency.









University of Glasgow

Conclusions

For a cluster with many fast worker nodes, like UKI-SCOTGRID-GLASGOW, it is important to have as fast a DPM head node as possible. IOwait on the database backend is a significant limiter of total performance, and reducing this significantly affects the total performance. Ultimately, however, you become limited by the capacity of your networking, and the seek rates on the disk servers themselves (as well as the overhead of GSI authentication per request). Initial testing at the other ScotGrid sites give the same early results as at Glasgow, and we are working to apply the lessons learned to their configuration.