PROCEEDINGS OF SCIENCE

Performance testing of SRM and FTS between Lightpath Connected Storage Elements

Brian GE Davies[1]
Lancaster University
Physics Department, Lancaster University, Lancaster, Lancashire, UK
E-mail: b.g.davies@lancaster.ac.uk

Roger WL Jones
Lancaster University
Physics Department, Lancaster University, Lancaster, Lancashire, UK
E-mail: Roger.Jones@cern.ch

We describe the configuration, testing and optimisation of file transfers with LCG middleware between the SRM storage systems at two LCG sites using a UKLight connection. We will also discuss recommendations for continued work.

Lighting the Blue Touchpaper for UK e-Science - Closing Conference of ESLEA Project
The George Hotel, Edinburgh, UK
26-28 March, 200

[1] Speaker

PoS(ESLEA)008

1. Introduction

ATLAS[1] is one of 4 large High Energy Physics (HEP) experiments physically based at CERN, Switzerland which will produce tens of PetaBytes of data each year. Since it is impractical to host and process this and associated Monte-Carlo data at a single site, high bandwidth data transfers between the hundreds of LHC Computing Grid[2] (LCG) sites around the world are needed. Within ESLEA[3], the ATLAS exploitation group, with help from the UK GRIDPP[4] community and in coordination with LCG service challenges aimed to establish the ability of the LCG middleware and hardware implementations to transfer large data volumes using the UKLight dedicated lightpath network.

ESLEA used in part UKLight to connect the LCG Tier1 centre at the Rutherford Appleton Laboratory (RAL) to Lancaster, which is a part of a distributed Tier2 (NORTHGRID) within the UK. This tier to tier connection crosses Regional Networks (RNs). The challenge of crossing these boundaries and the need for access to both production and research networks requires good communication between end-site system administrators, the regional network operators and the national network organisations.

2. Configuration

In order to optimise and analyse the use of the available bandwidth, an effective network, hardware and software configuration is needed. Configuration design should minimise obvious bottlenecks in performance. Since neither ESLEA nor the LCG are sole users of the RAL services, ESLEA worked within the RAL production framework to reduce interventions and carried out optimisations at Lancaster, where it has greater control and flexibility of hardware, software and network solutions.

2.1 Network configuration

Many factors affect the useful bandwidth on a production network. These include variable third party usage and bandwidth limitation, increased packet loss and jitter caused by multi-hop and variable routing and variable congestion of links. Dataset size and parallelisation of data streams were investigated. A private point to point link permitted complete control of the number of data flows and connections were allowed and so increased the ability to monitor rates between end-sites.

Dual homing of both hardware and software was initially considered. As this solution was not tested at the time, a network solution was used. By organising static routing and local network configuration, it was possible to allow both traffic flows across the dedicated lightpath for data transport whilst also allowing communication between internal and external services over the production network. One consequence of static routing is the need for all end-hosts' routing tables to be correctly configured and confirmed to ensure appropriate routing. The network configuration allowed an increased available nominal bandwidth from an intentional 100Mbps bottleneck (to avoid non-LCG site network congestion) over the production network

to 1Gbps over lightpath. We were also able to bypass a 400Mbps firewall. The lightpath also reduced router hops to two from twelve which leads to less potential for lost/reordered packets and associated network performance effects. With a round trip time (RTT) of 6.5ms and nominal bandwidth of 1Gbps, accepted standard TCP tuning techniques such as TCP window and memory buffer size optimisation of the native version of the linux-2.4 kernel to improve line usage were applied[5].

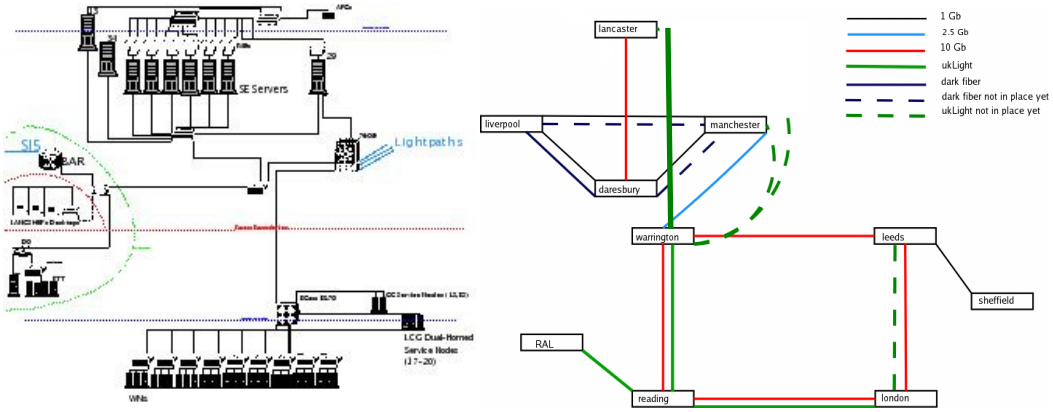


Figure 1- The Local Lancaster Network topology and the network topology connecting NORTHGRID Tier2 sites to RAL Tier1

2.2 Hardware/Software configuration

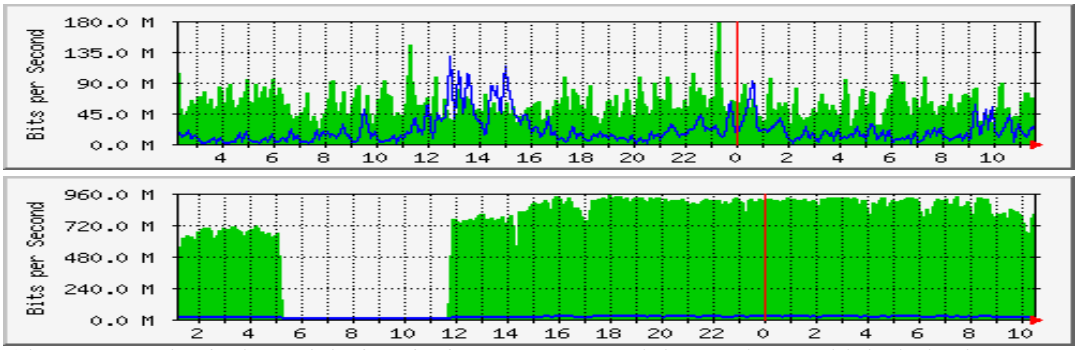
LCG file transfers use a storage resource manager (SRM) interface as a front end to an extended disk system. At Lancaster, we deployed a dCache[6] storage element (SE) installation onto a system consisting of a head node and six I/O servers, each with two 6TB RAID5 arrays. This allowed us to test both single and parallel concurrent transfers. CASTOR[7] (an alternative SE system) and dCache were both deployed and tested as the RAL end-point system. These systems had various numbers of file servers assigned to each endpoint throughout the testing procedure. In addition to an SE, several other LCG services and clients were installed to progress towards full distributed data management. Of particular importance were the File Transfer Service (FTS) and user interface (UI). The FTS, in conjunction with a UI allows file transfers from both disk-SRM and SRM-SRM. The SRM copies themselves are controlled by the dCache srmcp command. Transfers were initiated and controlled by two methods. The first method used a BASH command line script to initiate copies and deletions of files using loops and system sleeps. FTS managed transfers were controlled using the filetransfer.py script supplied by GRIDPP storage group. This incorporates another level of complexity of the software stack, as it requires extra communication with external servers leading to additional overhead. The FTS uses “channel management” to control gsi secure file transfers. This adds the ability to manipulate the number of concurrent streams and files transferred between two LCG sites, whilst channel status control enables complete transfer initialisation, cessation and retries. However, the FTS also increases the communication overhead of the transfer compared to a single SRM initiated command which in turn has its own overhead. The overhead from BASH onto srmcp is smaller than FTS and the filetransfer.py script. However long term functionality that FTS provides is needed for long term experimental use and so cannot be ignored.

2.3 Monitoring

Monitoring of rates, file storage and system diagnostics were achieved with various tools. MRTG and similar RRD tools were useful for both instantaneous rates and recording historical data of network traffic flows. Files copied using BASH scripts were checked with the commands ls and du. Python controlled FTS transfers were also capable of giving timing and failure rates. Both storage completions and rates werer cross-checked with Ganglia monitoring of SRM servers and SNMP walk information of routers.

3 RAL dCache to Lancaster transfers

For a single 1GB file transfer using srmcp an instantaneous rate of 330Mbps was achieved. However when incorporating the srmcp communication overhead, this rate dropped to an aggregate sustained rate of 195Mbps. Parallelisation of files transfers using a BASH script allowed a sustained rate of near line speed of over 900 Mbps with a peak rate of 946Mbps. This was accomplished with 20 concurrent file transfers from RAL to Lancaster. This rate also produced a back traffic rate of 18Mbps (2% of forward flow) which is presumed to be a summation of ACK packets inter-gridFTP door communication. Staggering of transfer initialisation also improved data rates by avoiding the concurrent dead time caused by concurrent initialisation/cessation of individual transfers. Further evidence of the effect of transfer dead time comes from studying the effect of file size on aggregate rates. The rate for a single file test between two particular servers increased from 150 to 180 Mbps with an increase from 1 to 10 GB file size. Sustained rates of 800 Mbps for 24 hours (Figure 2) and over 500 Mbps aggregated for a one week period were obtained. This corresponded to 8TBytes and 36TBytes of data transferred. Figure 2 also shows typical current production network rate usage.



(Figure 2- MRTG plot of 32 hour periods of transfers from RAL to Lancaster during normal usage and during load testing).

The drop from 800 to 500 Mbps between 24 hour and the week test was caused by authentication errors due to the user's grid certificate proxy expiring. Fail-over to the production network, caused by lightpath downtime, was successful in that no manual intervention was needed. The 100Mbps bottleneck imposed on the system led to full congestion of the production link with only a few concurrent transfers. This led to communication and timeout errors between FTS, UI and SRM services leading to dramatic drop in successful transfers. FTS controlled transfers for a single-stream, single-file transfer gave a rate similar to that of manual srmcp transfers (150Mbps). However, competition with production traffic using the FTS

channel led to an uncertain and unstable number of concurrent test files being transferred with FTS. Additional FTS server load from other experiments and end-sites caused lower transfer rates than from BASH script controlled transfers.

The modes and argument values of the dCache srmcp command and its effects needs to be studied. Variations in the direction of transfer and end-host initiation may explain directional rate variance. This observed change in rate may be an effect of the passive/active nature of the GridFTP or an effect of multiple/single stream transfers. It may also be a result of different SRM setups (such as pool balancing and file location.); or could also be an effect of disk I/O limitations of particular file-systems involved in the transfers.

3.1 RAL Castor to Lancaster Tests

Tests of the Castor system at RAL to Lancaster were successful but no extensive data loading rates are currently available. This confirms that the lightpath network topologies can allow multiple SEs to function at a single site. Initial rates obtained gave 600Mbps into Lancaster and 400Mbps out of Lancaster for single direction transfers. Rates of 200Mbps (in) and 300Mbps (out) for bi-directional tests with similar parallelisation were achieved. Initial tests of failure rates give a figure of 51 failed transfers from 851 1GB files transferred in a 12 hour period.

4 Future Plans

Following completion of the ESLEA project, we plan to continue testing of the CASTOR system at RAL. We intend to continue file transfers to the Netherlands over UKLight, connecting to a dCache system hosted by the LCG site at SARA. We hope to implement the UDT transport protocol into the LCG's Disk Pool Manager. The effects on data transport rates of additional LCG and ATLAS services, such as file catalogues and the ATLAS Distributed Data Manager software system (DDM) needs to studied. This work will be within the UK GRIDPP community and within LCG Service Challenge 4. An analysis of the effect of optimising the operating system (with particular focus on kernel version and automated TCP/IP tuning) might be studied in conjunction with the planned upgrade of the LINUX kernel version to 2.6.

References

[1] ATLAS homepage: http://atlas.web.cern.ch/Atlas/index.html

[2] LCG homepage: http://lcg.web.cern.ch

[3] ESLEA homepage: http://www.eslea.uklight.ac.uk

[4] GRIDPP homepage: http://www.gridpp.rl.ac.uk

[5] TCP-tuning: http://dsd.lbl.gov/TCP_Tuning

[6] dCache homepage: http://www.dcache.org

[7] CASTOR homepage: http://www.castor.org