Enabling advanced high performance networks and end-systems for Grid applications

Frank Saka[1], Nicola Pezzi[1], Andrea Di Donato[1], Javier Orellana[1], Peter Clarke[1], Yee-Ting Li[1], Stephen Dallison[3], Richard Hughes-Jones[3], Saleem Bhatti[1], Richard Smith[1], Robin Tasker[4]

(1) Networked Systems Centre of Excellence, University College London, Gower Street, London. WC1E 6BT.
(3) Department of Physics and Astronomy, University of Manchester, Manchester M13 9PL
(4) CCLRC, Daresbury Laboratory, Warrington Cheshire. WA4 4AD

Abstract

The MB-NG project brings together users, industry, equipment providers and e-science applications. The project aims are: to construct a high-performance leading edge quality of service (QoS) network; to demonstrate end-to-end managed bandwidth services in a multi-domain environment and to investigate high performance data transport mechanisms for Grid data transfer across heterogeneous networks. We report on the major successes in the area of QoS and managed bandwidth, the achievements in the area of end-hosts and the benefits to applications.

1 Introduction

The MB-NG project is a major collaboration between different groups. This is one of the first projects to bring together users, industry, equipment providers and leading edge e-science applications. The project aims are: to construct a high-performance leading edge quality of service (QoS) network; to demonstrate end-to-end managed bandwidth services in a multi-domain environment in the context of Grid project requirements and to investigate high performance data transport mechanisms for Grid data transfer across heterogeneous networks.

Now in the last of its two year program, MB-NG has demonstrated major successes in the area of QoS and managed bandwidth; end host configuration; high bandwidth data transfers in collaboration with DataTAG; and the benefits to applications.

This paper contains an overview of the work carried out and references on where to find more details on each topic.

2 The network

The MB-NG network is shown in Figure 1. It is composed of a core network surrounded by three edge networks at Manchester, University College London (UCL) and the Rutherford Appleton Laboratory (RAL).

The core of the network consists of Cisco 12000 Gigabit Switching Routers (GSR) connected at 2.5 Gbit/s. The edge networks are made up of Cisco 7600 Optical Switching Routers (OSR) interconnected at 1Gbit/s.

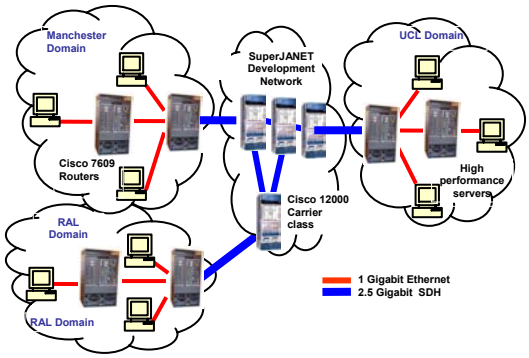

Figure 1. The MB-NG network

2.1 QoS

The Quality of Service (QoS) deployed in the MB-NG network is the differentiated services (diffserv) model [diffserv] where IP packets are marked with a diffserv codepoint (DSCP). Packets receive preferential treatment in the network based on this value.

An example of the end-to-end experiment performed had traffic being sent from the UCL domain to Manchester and RAL at 3 Gbit/s. This caused congestion on the egress interface of the UCL 7600 connecting to the core network. When a “priority” TCP flow was sent from London to Manchester, the throughput it achieved was near zero as the background traffic used up the available bandwidth of 2.5 Gbit/s. At around 185 seconds into the experiment, QoS was configured in the network, reserving 1 Gbit/s for the priority flow. The result is that the priority flow is protected from the background traffic as shown in Figure 2. This deployment is the first example of a long

distance network running QoS at 2.5 Gbit/s in UK. It required leading edge Cisco interface cards.

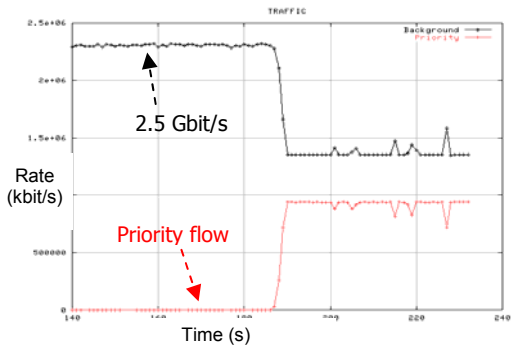

Figure 2. QoS in action. During the experiment, QoS is enabled and the priority flow is protected from the background traffic.

Extensive work on QoS using multiple classes has been done on network equipment in standalone mode and through the MB-NG, DataTAG and GEANT networks [eunice] [addqos] [infn].

2.2 MPLS and managed bandwidth service

One of the aims of MB-NG is to investigate the provisioning of a managed bandwidth service (i.e. a guaranteed bandwidth allocation for a particular flow) over IP networks. Multi-Protocol Label Switching (MPLS) was deployed in the MB-NG network to test its Traffic Engineering (TE) capabilities and to prove what can be done in terms of the provisioning of a managed bandwidth service.

MPLS traffic engineering (MPLS-TE) allows engineering of the route taken by packets by switching on label number.

To create tunnels between end points many different protocols need to be enabled in the network, including:

- MPLS with TE extensions;
- A link-state Interior Gateway Protocol (IGP) with TE extensions. This is required because the path followed by the tunnel is calculated at the head end router. So this must have a complete map of all the network resources. We used ISIS (Intermediate system to intermediate system) as the IGP.
- RSVP (Resource Reservation protocol) with TE extensions. This is used as the label distribution protocol (LDP) to carry information about resource availability on each link.

Regarding the managed bandwidth it is important to clarify that there is a difference between forwarding plane and control plane. MPLS-TE will "guarantee" bandwidth on the control plane by using RSVP signalling to create the tunnel path. However, by default there is no correlation between the MPLS TE control plane and the forwarding plane and without this correlation, there is no guaranteed bandwidth in the network.

To implement a guaranteed bandwidth service, the correct associations between the control plane and the forwarding plane need to be created. In particular:

- Traffic entering each tunnel interface should be policed to the tunnels configured bandwidth such that the perceived service quality does not deteriorate.
- Traffic entering each tunnel interface should be marked to a unique IP precedence (or DSCP if DiffServ is being used) to indicate that it is tunnel traffic. By default, the three most significant bits of the DSCP field is copied into the Experimental field (EXP) of the MPLS header.
- Different queuing strategies should be configured on each interface that belongs to the MPLS cloud.

MPLS was configured and demonstrated in the core of the network. Used together with QoS, MPLS-TE provides a solution to the managed bandwidth problem. It can be used to "carve out" protected pipes for traffic flows. More details on the exact use and example configuration can be found at [mpls]

3 Middleware

We deployed the middleware software called grid resource scheduling (GRS) [grs] to micro-managed capacity allocations at the edge sites (a la diffserv). GRS currently assumes an overprovisioned core.

Figure 3 shows the GRS architecture. The Network Resource Scheduling Entities (NRSEs) at each end-site (each domain) are responsible for receiving user requests for protected capacity, checking if the request can be honoured site-to-site, and then issuing configuring instructions to the local network elements in order to provide the requested QoS. The arrows show the paths for signalling messages. Host end-systems may send requests to the NRSEs for a protected QoS allocation, in the form of a local service level agreement (SLA) (1). NRSEs then communicate to arrange

"booking" of the request at both ends of the communication path (2). At the appropriate times, the NRSEs in each domain issue local instructions to programme the requested QoS into the network elements. The signalling protocols (1) and (2) are application-level protocols, whilst (3) is a purely local protocol.

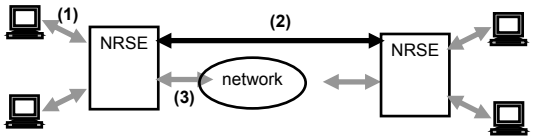

Figure 3. The GRS architecture.

MB-NG is the first successful deployment and use of GRS in a WAN [grs]

4 High performance data transport

Newly emerging TCP stacks have been shown to have highly increased bandwidth utilisation over standard TCP in long delay high bandwidth and multi-user network environments [hstcp] [scalable] [htcp]. This allows a single stream of a modified TCP stack to transmit at rates that would otherwise require multiple streams of standard TCP. These performance parameters have been investigated together with end-host performance behaviour such as the effects of PCI bus interaction, disk behaviour and various RAID solutions.

4.1 Protocol performance

Figure 4 shows how the performance of the new stacks compare with standard TCP on the DataTAG network. This shows that the performance of standard TCP is fairly poor irrespective of the background rate. The newer TCP stacks show better responsiveness to the background rate.

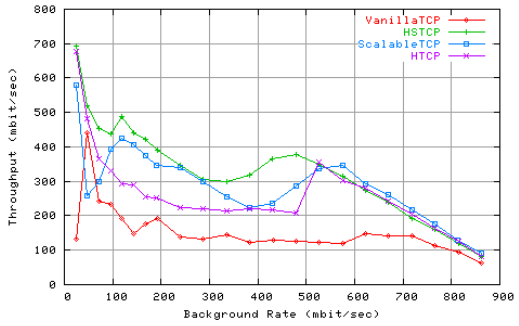

Figure 4. The performance of various TCP stacks on the DataTAG network. RTT~120ms.

On the MB-NG network, the performance of all the protocols was similar. This is due to the much smaller round-trip time (RTT) of around 6 ms compared to DataTAG's of around 120 ms. Extensive tests on how these protocols perform under various network conditions including issues of fairness and interoperability with QoS are presented in [yeepfldnet], [pfldnet] and [terena].

4.2 Disk-to-disk performance

The scientific community is mainly interested in bulk data transfers. Numerous hardware configurations with raid arrays has been tested to assess Disk-to-disk performance [yeepfldnet].

Figure 5 shows the write speeds for various raid controllers in a raid 5 configuration (4 maxtor 160Gbytes disks with 2Mbytes cache 7200 rpm) and how they compare with a normal system disk.

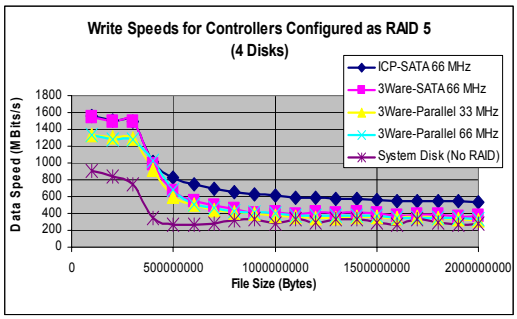

Figure 5. The write speed for various controllers in raid5 configuration.

We see that a normal system disk achieves a rate of 300 Mbit/s for large files (>20 Gbyte) whereas the best achieved with a raid array is 600 Mbit/s. For reading, the maximum rate measured was 1300 Mbit/s.

5 Application performance

In a Collaboration between the BaBar particle physics experiment and MB-NG we demonstrated high performance data transport using novel variations on the TCP/IP transport protocol. Data from the BaBar experiment stored at RAL is transferred to Manchester on a regular basis. The transfers are large (typically about 800 GBytes to 1 TByte) and currently take at least 60 hours over the SuperJanet-4 production network. An example showing this type of transfer is shown in Figure 6. This shows the transfer of 40 Gbytes of BaBar data which takes from 12:00 to 16:00.

Figure 7 shows the transfer of 900 Gbytes through the MB-NG network between 18:00 and 02:00. This transfer takes less than 10 hours

compared to around 60 hours through the production network.

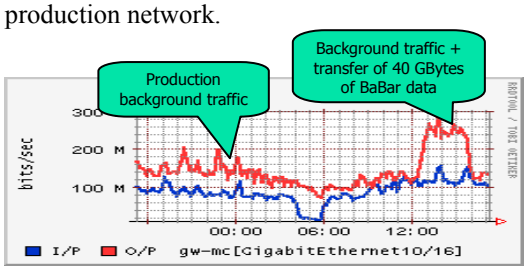

Figure 6. The transfer of 40 Gbytes of BaBar data through the production network.

Work with applications are ongoing. The results will be made publicly available on the MB-NG website [mbng].

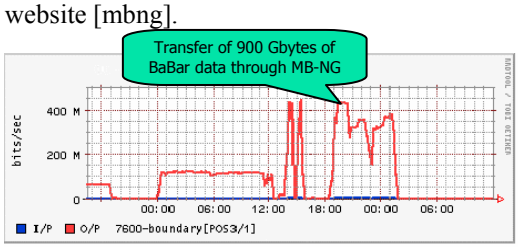

Figure 7. The transfer of 900 Gbytes of BaBar data through MB-NG

6 Conclusion

The major successes of the MB-NG project in the area of QoS and managed bandwidth are: enabled a leading edge UK diffserv enabled network running at 2.5 Gbit/s; configured and demonstrated the use of multi-protocol label switching (MPLS) traffic engineering to provide tunnels for preferential traffic and in addition deployed a middleware to dynamically reserve and manage the available bandwidth on a per-flow level at the edges of the network.

The achievements in the area of end-hosts are: assessed disk-to-disk performance; investigated high-bandwidth data transfers within the UK and as a result of our successful collaboration with the DataTAG project, across a transatlantic link offering orders of magnitude larger round-trip time. Finally, we demonstrated the benefits this type of advanced network environment brings to the scientific community compared to today's "production" network.

Acknowledgements

We would like to thank our project partners: UKERNA, Cisco, Spirent, University of Southampton, Cambridge University and Lancaster University.

References

[addqos] CERN, INFN, INRIA, PPARC, UvA "High performance networking: End-to-end Interdomain QoS" https://edms.cern.ch/file/431720/5/WP2-D2.3-Final-R2.pdf May 2004.

[diffserv] S. Blake et al "An Architecture for Differentiated Services" RFC 2475 http://www.ietf.org/rfc/rfc2475.txt Dec. 1998

[eunice] J. Orellana, A. Di Donato, F. Saka P. Clarke "Benchmarking QoS on Router Interfaces of Gigabit Speeds and Beyond". 9th Open European Summer School and IFIP Workshop on Next Generation Networks (EUNICE 2003), Budapest-Balatonfured, Hungary, September 2003

[grs] R. Smith, S. Bhatti "Network resource scheduling" http://www.cs.ucl.ac.uk/staff/S.Bhatti/grs/docs.html 2004

[hstcp] S. Floyd, "High Speed TCP for Large Congestion Windows", IETF Internet Drafts, http://www.ietf.org/internet-drafts/draft-ietftsvwg-highspeed-01.txt

[htcp] D. Leith, R. Shorten, "H-TCP: TCP for high Speed and Long Distance Networks" PFLDNet-2 Argonne, Illinois USA Feb. 2004

[infn] A. Di Donato "Description, results, plots and conclusions about LBE tests between INFN and UCL: case congestion at the edge only"http://www.mb-ng.net/technical/reports.html Feb. 2003

[mbng] MB-NG: Managed Bandwidth - Next Generation http://www.mbng.net/

[mpls] N. Pezzi "Deployment of MPLS over MB-NG" http://www.mb-ng.net/technical/reports.html June 2004

[pfldnet] A. Di Donato, Y. Li, F. Saka and P. Clarke "Using QoS for High Throughput TCP Transport Over Fat Long Pipes". PFLDNet-2. ANL, Argonne, Illinois USA. Feb. 2004

[scalable] T. Kelly, "ScalableTCP: Improving Performance in High Speed Wide Area Networks", Proceedings of PFLDNet-1, Feb. 2003

[terena] A. Di Donato, F. Saka, J. Orellana and P. Clarke "On the joint use of new TCP proposals and IP-QoS on high Bandwidth-RTT product paths" TERENA networking conference 2004. 7-10 June 2004. Rhodes, Greece.

[yeepfldnet] Y. Li, S. Dallison, R. Hughes-Jones, P. Clarke "A Systematic Analysis of TCP Performance" PFLDnet2004. Argonne National Laboratory Argonne, Illinois USA February 2004