



MB-NG Project

First year report

-10 June 2003-

Document identifier:	MB-NG-Doc-T0-v1.2-1stYearReport
Relevant Task	Task 0
Date:	10/06/2003
Version:	Version 1.2
Document status:	Final
Description	First year report on in the MB-NG project
Contact:	F.Saka fs@hep.ucl.ac.uk

Table of Contents:

1	Structure of document.....	4
2	Overview.....	4
2.1	Summary of the first year	5
2.1.1	The Network	6
2.1.2	Early QoS tests.....	7
2.1.3	Gigabit per second Data Transport	8
2.2	Summary of plans for the second year.....	8
3	Formal EPSRC Report.....	10
3.1	EPSRC Template	10
3.1.1	Resources – staffing, DTAs, equipment etc.....	10
3.1.2	Collaboration – user engagement (existing collaborators, new collaborators), international, other e-Science projects, other Science projects, within Academic team)	11
3.1.3	Dependencies on other projects/Dependencies of other projects on Pilot – details of dependencies	12
3.1.4	Progress against Milestones – prototypes etc	12
3.1.5	IPR issues – collaborative agreements, problems encountered, solutions	13
3.1.6	Technical and Published Outputs to Date – open source/standards software, patents, papers, highlights, others (please specify)	13
3.1.7	Technical and Published Outputs Planned – open source/standards software, patents, papers, others (please specify)	13
3.1.8	Training Issues	13
3.1.9	Management Procedures	14
3.1.10	Relationship within Global e-Science Community.....	14
3.1.11	Other issues	14
3.2	Original Deliverables from the project proposal	14
4	QoS Network and Measurements	18
4.1	QoS: the network engineering solution in building a multimedia IP network.	18
4.2	Standalone tests on the Cisco12000 series router	20
4.3	End-to-end tests: CORE bottleneck test results	22
4.3.1	Virtual bottleneck: QoS_On_Shape.....	23
4.3.2	Physical bottleneck: OC3.....	27
4.4	End-to-end tests: EDGE bottleneck	29
4.5	QoS - summary	31
4.6	Future work	31
	References.....	31
5	Middleware	32
5.1	Overview.....	32
5.2	Middleware Survey	32
5.3	GARA Status	33
5.4	Current Testbed.....	33
5.5	Future Activities.....	34
6	High throughput program	35

6.1	High Throughput Data Transfers	35
6.2	What's Involved in Achieving High Performance Networking	35
6.3	Some Issues Affecting Performance	36
6.3.1	Investigation of Packet Loss	36
6.3.2	Investigation of Errored Packets	37
6.3.3	The importance of the Transmit Queue length	38
6.4	End-to-End Memory-memory Tests with UDP	39
6.5	End-to-End Memory-memory Tests with TCP	41
6.5.1	24 Hour Throughput with a Single TCP Stream	41
6.5.2	Multiple TCP Streams to one Remote Host	42
6.6	GridFTP Data Transfers	45
6.6.1	Throughput Measurements	45
6.6.2	Analysis of the Transfers	46
6.7	http Data Transfers	47
6.8	Transfer of Radio Astronomy VLBI Data	49
6.9	Conclusions	49
	References	51

Part 1: Overview and Formal EPSRC Report

Progress Report on EPSRC e-Science Pilot Project

Project: MB-NG

PI: Professor Peter Clarke

Date: 23/5/03.

1 Structure of document

This document is submitted by MB-NG to the EPSRC as a required report following the first year of activities.

The document is structured in the following way:

- In section 2 we give an overview of the objectives of the MB-NG project and the progress made towards them in the first year. We highlight the principal successes and give some illustrations (which could be used in further summary reports or presentations by EPSRC/PPARC). We then list our objectives for the second year.
- In section 3 we give a formal report based upon the template supplied by EPSRC. This should be read in conjunction with, section 2.
- In all of the following sections we present the work carried out in the first year in full detail. We believe this to be an impressive record, representing a very substantial amount of progress made in many technical areas.

2 Overview

The MB-NG project has reached the end of its first year of operation. This document is a report on the achievements made within this year and the goals for the next year.

The primary high level objectives of MB-NG are:

- **To demonstrate leading edge end-to-end managed bandwidth services in a multi-domain environment, in the context of Grid project requirements.**
- **To demonstrate very high performance data transport mechanisms in the context of Grid projects.**

These objectives define the value added which MB-NG brings over simpler more limited environments. MB-NG will demonstrate the benefits of leading edge network services and high performance data transport to real e-Science applications and hence show the path to new “ways of working” needed for the evolution of Grids.

Our more detailed objectives, which follow from the above, are:

- i) to construct the first very high performance – leading edge QoS enabled network in the UK.
- ii) to demonstrate managed bandwidth (MB) across multiple administrative domains simulating a Core network and local site networks, addressing SLA/SLS and policy issues.
- iii) to demonstrate use of MB services by Grid applications.
- iv) to interface MB services to Grid middleware APIs.
- v) to pilot new protocols to enable very high performance network links to be utilised by high demand Grid applications, i.e. to demonstrate gigabit per second (Gbit/s) scale flows.
- vi) to provide a vehicle for UKERNA to gain experience of the use of MPLS for traffic engineering in the Core.
- vii) to pilot services in an international environment in collaboration with EU and US groups.

These detailed objectives are derived directly from the tasks specified in the original proposal, although the detailed schedule has changed to take advantage of opportunities, and as a result of a longer than expected equipment delivery time.

2.1 Summary of the first year

In the first year we have achieved the following successes:

- We have constructed the first leading edge QoS enabled network in the UK comprising 4 independent domains linking Manchester, London and CLRC-RAL, and permitting different classes of service to be handled at up to 2.5 Gbit/s.

- We have made a first demonstration of end-to-end differentiated packet services upon this network
- We have performed sustained data transport across the network at speeds near 1 Gbit/s for periods in excess of 24 hours.
- We have performed international differentiated services demonstrations through MB-NG's participation with:
 - o The DataTAG project and
 - o GÉANT
- We have already begun to collaborate with international groups to deploy Grid oriented network resource allocation middleware upon MB-NG,

We have thus achieved (i) completely and (ii) in part and we have and made major strides towards (iv), (v) and (vii). We illustrate these achievements briefly in the rest of this section, and give a much more detailed account in later sections.

2.1.1 The Network

Figure 1 shows the testbed network. This comprises three “edge” domains built from CISCO 7600 series Optical Switch Routers, with the most advanced interface cards (2.5 Gbit/s SDH and 1 Gbit/s Ethernet Wide area network QoS enabled line cards). These edge domains are connected via the SuperJANET development core network comprising 4 carrier class CISCO GSR 12000 series routers similarly equipped with leading edge 2.5 Gbit/s QoS enabled line cards. The network connects to “application” sites in the UCL and Manchester campuses, and CLRC–RAL

In constructing this network we have carried out very detailed characterisation tests of each component. Not only has this provided a very firm foundation for understanding the end-to-end performance, but it allows us to disseminate this important information to the UK education and research network community. Thus one of the very important “by products” of MB-NG is a series of Technical Information Documents which we will publish to the UK research and education community. The first of these is already public and can be seen at <http://www.mb-ng.net/technical/index.html>.

In addition MB-NG gave a detailed technical presentation at “NetworkShop 31” held in York University in April 2003. This is the premiere meetings in the UK, attracting engineers and decision makers from all UK HE and FE institutes, MAN providers, and industry.

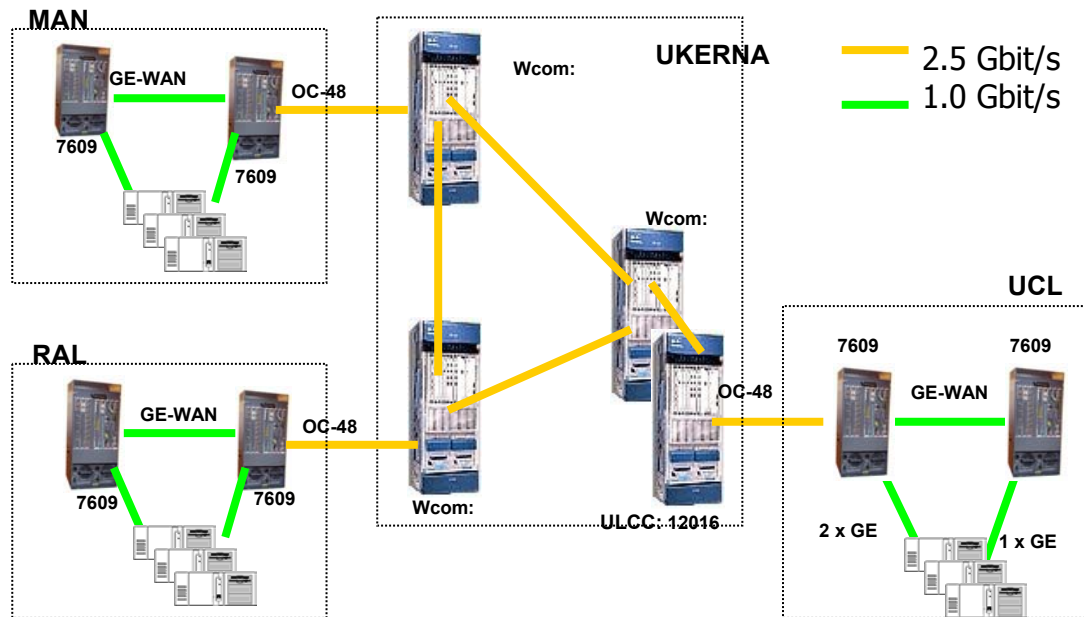


Figure 1. The MB-NG testbed network.

2.1.2 Early QoS tests

Figure 2 illustrates the first QoS demonstration carried out upon this network. Here we see the effect when two classes of traffic are presented to the network, so called “best effort” (BE) IP and “less than best effort IP” (LBE). As we proceed from left to right we see both classes taking an increasing amount of bandwidth as their rates are increased. Then, as the link capacity is reached, the LBE class is rejected in favour of the BE class which continue to increase. Simple though this demonstration is, it provides a proof of principle for the idea of sending massive data sets which are not time critical via the LBE class as part of systematic data management. Such replication can then safely take all spare link capacity without fear of causing other more time critical applications to be harmed – for whenever there is a significant ordinary traffic load the data replication backs off. This is the network equivalent of “idle CPU cycle stealing”.

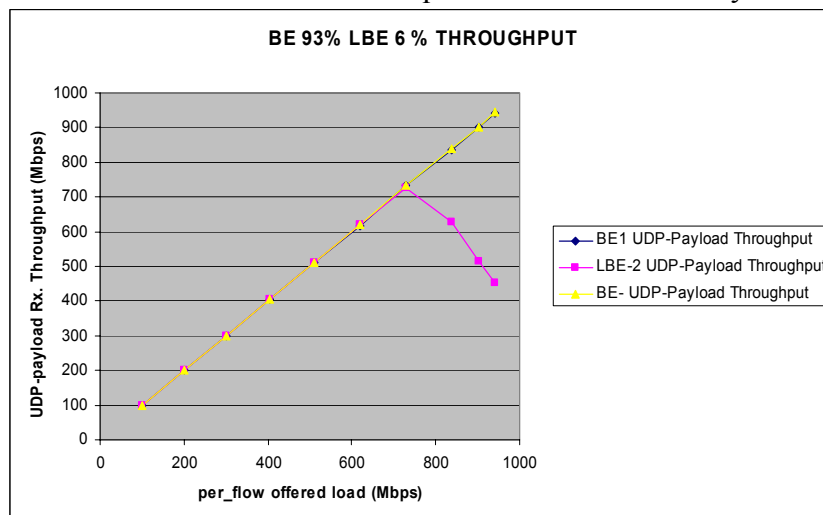


Figure 2. An example of differentiation between Best-Effort and Less-Than-Best-Effort traffic.

2.1.3 Gigabit per second Data Transport

It is well known that today applications cannot easily make use of the high capacity networks we have available. This is known as the “Wizard Gap”, alluding to the fact that “out of the box applications” achieve typically 20-100 Mbit/s on the wide area network, when this network itself has up to 10 Gbit/s capacity – only network Wizards can bridge the gap. The reasons for this are well known and include poor TCP configuration settings in standard deployments, the inadequacies of TCP itself to work properly at very high rates, failings of the end host interface cards and finally under-performant disk systems. The MB-NG project is working with international partners to bring this knowledge to the applications, and as a first step along the way to this, Figure 3 demonstrates a sustained TCP connection at nearly 1 Gbit/s using HighSpeed TCP (HSTCP), one of the new proposed TCP implementations.

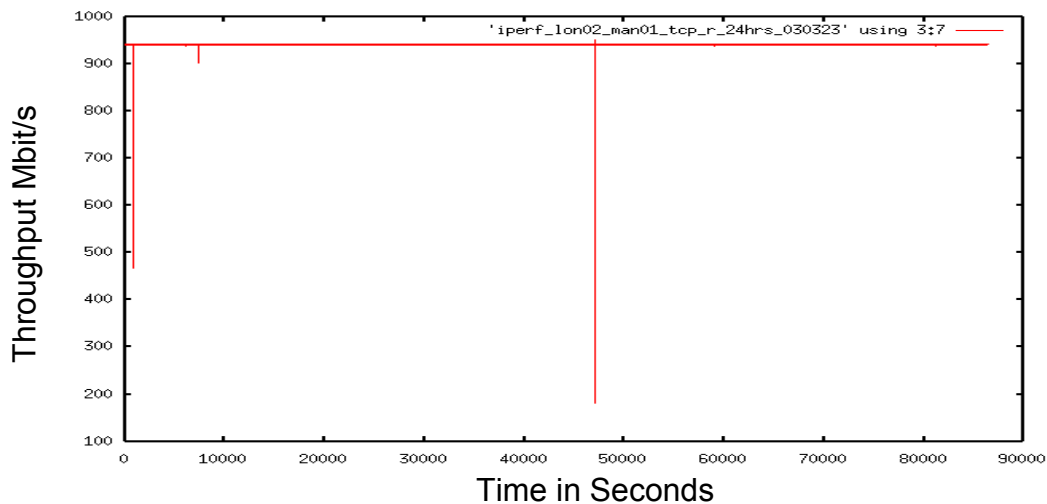


Figure 3. 24 Hour time series throughput for memory-memory transfers with 1 stream using the HighSpeed TCP stack.

2.2 Summary of plans for the second year

In the second year of the project, MB-NG will move from what has primarily been a construction and characterisation phase – to an exploitation and demonstration phase. In so doing we will realise the value added of this project over and above equipment tests. Specifically our objectives for the second year will be:

- 1) To demonstrate the benefit of “multi service networks” to
 - a. Particle Physics data transport
 - b. The Reality Grid Core e-Science project

- 2) To demonstrate the use of advanced data transport protocols to real experiments by transporting real experiment data between sites at speeds approaching Gbit/s
- 3) To deploy and demonstrate simple end-to-end resource allocation middleware. Specifically we intend to show a simple Grid application (e.g. GridFTP) gaining access to a network service through a suitable middleware API.
- 4) To develop a simple managed bandwidth interface to users and show how this can be provisioned using the underlying QoS enabled network
- 5) To carry different classes of service using MPLS
- 6) To link MB-NG to UKLIGHT and through this, achieve the extended tests to the EU and USA.

3 Formal EPSRC Report

In this section we address the report formal request form the EPSRC and the milestones in the original proposal.

3.1 EPSRC Template

3.1.1 Resources – staffing, DTAs, equipment etc.

Staffing

MB-NG has three full time grant supported staff in post (as provisioned in the grant)

Dr. S. Dallinson
Mr. N. Pezzi
Dr. F. Saka

Mr. Pezzi replaced Mr. N. Simpson who started with the project initially, but who left at approximately 6 months due to financial difficulties. This introduced a few months staffing gap, but this has not caused serious problems to the project. We expect to achieve the stated goals by then end of the “grant break” period.

The project uses engineering time from Mr. I. Bridge (UCL), Dr. C. Seelig (RAL) and Mr. A. Ryan (Manchester).

CISCO has assigned 25 days of a network design engineer to work specifically with the project.

In addition the project leverages strongly the DataTAG project. Work going on in DataTAG, which is being carried out by three engineers at UCL working side by side with MB-NG staff, is directly relevant to, and being deployed upon, MB-NG. This particularly applies to the Middleware thread and policy and configuration issues.

Equipment

The main project equipment costs are summarised in the table below:

Item	Cost inc vat
Network Routing Equipment ¹	£227,000
Test and measurement equipment	£26,000
Misc (notebooks, connectors, ..)	£6,000
RAL link (commit)	£50,000
Total	£309

¹Routing equipment with a list price value in excess of £1M was supplied by CISCO

The first year has seen the construction of the testbed network, which naturally represents the majority of the required equipment.

The project is on budget and no problems are currently foreseen.

3.1.2 Collaboration – user engagement (existing collaborators, new collaborators), international, other e-Science projects, other Science projects, within Academic team)

The collaborations achieved as a result of MB-NG are many-fold:

- Amongst existing collaborators: UKERNA and CISCO staff have been deeply involved in the project from the beginning. Specifically Mr. J. Sharp manages the UKERNA collaboration, and Dr. D. Salmon and Mr. M. Allenby are involved on a regular basis. Several CISCO engineers have been involved with respect to technical problems, and now we have an assigned project engineer.
- The DataTAG project is an EU funded Grid project to pilot QOS and high performance data transport on a 2.5 Gbit/s transatlantic link. MB-NG management staff are prominent in this project (P. Clarke is the UK “PI”, R. Tasker of CLRC DL is the relevant workpackage manager). DataTAG has resulted in three (3) network engineers paid by the EU and working in UCL alongside the MB-NG team. The goals of the projects are aligned and hence the synergy is complete. This has provided an enormous benefit to both projects. DataTAG recently had a successful EU review.
- We have undertaken collaborative work with DANTE (GÉANT) on several occasions:
 - We mounted a joint demonstration with DANTE at the framework 6 launch in Brussels in Oct 2002
 - We mounted a demonstration at the IGrid2003 conference in Amsterdam in September 2003.
 - We have undertaken tests of LBE traffic class across the GÉANT backbone to Italy (details given below)
 - We have participated in the production of a promotional DVD for GÉANT.
- We have recently agreed to use the MB-NG project to perform a demonstration in collaboration with Reality Grid, a core eScience project.

- We collaborate with Zentralinstitut für Angewandte Mathematik in Germany and the University of Amsterdam in respect of deployment and testing of Grid network resource allocation middleware.
- Three of us have roles in the GGF in the network sector: P. Clarke is director of the “Data Area” which encompasses most network oriented groups. R Hughes-Jones is co-chair of the network performance measurements group, Y .Li is secretary of the high performance transport group.

3.1.3 Dependencies on other projects/Dependencies of other projects on Pilot – details of dependencies

- For the majority of its goals the project does not depend strongly on other projects to achieve the majority of its objectives.
- For the demonstration of the benefit to a non-HEP application the project depends upon Reality Grid
- For extensions of tests to the US the project depends upon the commissioning of UKLIGHT

3.1.4 Progress against Milestones – prototypes etc

The full progress against expanded milestones is described in the overview. In this, we set the achievements in context. The overview should therefore be read before reading this section. The primary output of the first year has been the construction of the first leading edge QoS enabled network in the education and research sector capable of handling differentiated services at up to 2.5 Gbit/s, and first demonstrations of sustained TCP flows at near 1 Gbit/s.

Formal progress against the milestones in the original proposal is given in the following subsection. Probably the only significant variation is that we have pushed the use of MPLS into the second year.

In summary we have achieved the majority of the objectives which were foreseen for the end of the first year, taking due consideration of the changing landscape and opportunities.

3.1.5 IPR issues – collaborative agreements, problems encountered, solutions

There are no fundamental IPR issues.

3.1.6 Technical and Published Outputs to Date – open source/standards software, patents, papers, highlights, others (please specify)

We have already given a talk on MB-NG at Networkshop 31 in York, April 2003. This workshop is aimed at network managers and technical staff from the JANET community. It provides a forum for discussion on the latest network technologies and brings together expertise from all the fields of networking.
<http://www.ja.net/conferences/networkshop/>

We are publishing a series of documents on QoS configurations on various routers and TCP/IP settings for high performance. These are designed to disseminate knowledge to the UK research and education community. The first of these can be found at
<http://www.mb-ng.net/technical/index.html>.

3.1.7 Technical and Published Outputs Planned – open source/standards software, patents, papers, others (please specify)

As mentioned above, we will be publishing many more technical information documents to the UK research and education community.

We have submitted a paper to EUNICE 2003 conference on Next Generation Networks, Hungry Budapest 8th-10th September 2003. The title of the paper is “Benchmarking QoS on Router Interfaces of Gigabit Speeds and Beyond”

We have submitted two talks (on QoS and high-throughput protocols) to the e-Science all hands meeting in Nottingham, 2nd-4th September 2003.
<http://www.rcuk.ac.uk/escience/documents/allhands2003webpage.htm>

3.1.8 Training Issues

There are no training issues

3.1.9 Management Procedures

The project is managed using the following procedures

- A series of documents sets out the high level goals and the experiments which encapsulate success
- From the above a detailed task definition and experiments definition document has been produced.
- A schedule chart (based upon Microsoft Project) is maintained which sets out the time schedule of each task and experiment.
- An “elements map” is maintained which charts the completion status of each element required to achieve the goals of an experiments.
- A regular bi-weekly operational meeting takes place. At this the Project chart and elements map are reviewed in detail against the planned schedule. Actions are constituted as appropriate.
- At approximately four monthly intervals a wider “collaboration meeting” is held.

All of these documents and tools can be viewed at the following URLs:

<http://www.mb-ng.net/management/>

<http://www.mb-ng.net/documents.html>

3.1.10 Relationship within Global e-Science Community

The relationship with the global e-Science community is described in section 3.1.2 (collaborations).

3.1.11 Other issues

There are no other issues at this time.

3.2 Original Deliverables from the project proposal

In this sub-section we comment specifically (but briefly) upon the formal milestones given in the original proposal. However we emphasise that, as described in the overview, MB-NG has substantially widened its goals to include proof-of-concept demonstrations with real applications, and the deployment of network resource access middleware – these were opportunities not foreseen at the time of the proposal and therefore do not appear in these milestones.

Task TM1: To understand the use of MPLS as a traffic engineering tool within the CORE SJDN.

Deliverables:

- Month 3: Procurement and installation of equipment
- Month 6: Initial demonstration of sustained throughput for different traffic classes.
- Month 9: Completion of work.
- Month 12: Final report

Status:

We have not started the MPLS tasks yet as it became apparent that this would be best performed in the second year.

The construction and characterization of the testbed took longer than expected due to (i) delays in delivery of equipment from CISCO (approx 4 months) but more importantly (ii) the leading edge nature of the interface cards as MB-NG required the most advanced versions which have only just become available.

We took the decision to concentrate upon achieving the pure IP based QoS goals first, and will then move to MPLS in the second year.

Task TM2: To demonstrate end-to-end traffic management across multiple domains using live Grid traffic.

Deliverables:

- Month 12: Initial demonstration of end-to-end guaranteed bandwidth and QoS. Interim report. Presentation of results at network venues.
- Month 18: Advanced demonstration including use of other QoS techniques.
- Month 24: Final report. Presentation of results at networking venues.

Status:

The month 12 deliverable has been achieved. We have (i) demonstrated QoS in the MB-NG network (ii) presented a technical report at NetworkShop 2003 and (iii) produced the first of a series of Technical Information Documents.

Task: TM3: To demonstrate end-to-end QoS and traffic management between the UK and USA.

Deliverables:

Deliverable dates must necessarily be less concrete at present, until the availability of a suitable transatlantic connection is established.

- Month 12: Interim report on progress and tests made to date.

- Month 24: Final report.

Status:

We have partially achieved the month 12 deliverable in that we have demonstrated QoS on the link from CERN to Chicago as part of DataTAG. However since a suitable direct link from MB-NG to the USA is only just becoming available via UKLIGHT we have not made the demonstration from the UK itself. This is in accord with the variability of suitable links as noted in the proposal.

Task: TM4: *To demonstrate end-to-end QoS and traffic management to CERN*

Deliverables:

Deliverable dates must necessarily be less concrete at present, until the availability of suitable connection is established.

- Month 12: Interim report on progress and tests made to date.
- Month 24: Final report

Status:

We have essentially achieved the month 12 deliverable through our collaborative work with DANTE/GÉANT (described below). In this work we piloted “Less than Best Efforts” tests between London and Italy using the GÉANT network. This entailed close cooperation with the DANTE engineers.

Task: TP1: *Demonstrate high performance transport applications across the WAN in a live Grid context, with a target of 1 Gbit/s.*

Deliverables:

- Month 9: Demonstration of reliable transport at > 100 Mbit/s over WAN.
- Month 18: Demonstration of reliable transport at > 1 Gbit/s .
- Month 24 Final report

Status:

We have exceeded the month 9 deliverable in several ways. We (i) have sustained 1 Gbit/s memory to memory, (ii) achieved 4-700 Mbit/s disk to disk (iii) been responsible for several high profile demonstrations at key venues. These were (a) the Framework 6 launch in Brussels, where we worked with DANTE as part of the EU IST stand to show QoS and high rate transport of VLBI data between Manchester and the Netherlands (b) a similar demonstration at the IGRID 2002 meeting held in the Netherlands and (iii) our

contribution to the PFLD high performance transport workshop
(<http://datatag.web.cern.ch/datatag/pfldnet2003/index.html>).

Part 2: Detailed description of Work and achievements

4 QoS Network and Measurements

4.1 QoS: the network engineering solution in building a multimedia IP network.

The traditional Best effort networks provide no means to differentiate between packets belonging to different applications. Typically, the routers of which these networks are composed have unpredictable temporal response to transient congestion and an inability to provide priority service to different classes of traffic.

The goal in a QoS enabled environment is to enable predictable service delivery to certain classes or types of traffic regardless of what other traffic is flowing through the network at any given time, that is, to create a multi-service IP network solution where traditional bursty traffic may share the same infrastructure as traffic with more rigorous latency, jitter, bandwidth and/or packet loss requirements.

The critical factor in the provision of QoS is how routers treat different types of packets during periods of transient and long term congestion. QoS capable routers must be based on an architecture that allows sufficient differentiated queuing and scheduling to be applied at each hop. They must be able to classify packets into different classes, queuing each class separately and assigning unique levels of processing priority to each class.

A traditional router chiefly focuses on **where** to send packets, routers for QoS-enabled IP networks must also control of **when** to send packets.

Network designers face a trade-off between the number of traffic classes carried by their networks and the number of traffic classes that their router's Classify-Queuing-Scheduling (CQS) architecture can handle. Some solutions are based on distributed edge-and-core architectures, where the cores are fast routers with limited CQS capabilities and the edges are slower but with more advanced CQS capabilities. The IETF Differentiated service (Diffserv) model is an example of such a solution and is the approach being followed in the MB-NG project.

The strength of a project like MB-NG is that the tests are carried out in a real geographically separated multi-domain network. Thus the project results have technical output that can be potentially deployed immediately in a real environment.

Furthermore, the MB-NG project is not just about implementing QoS over IP in a real network but also about implementing it in the most technologically advanced QoS-capable IP network possible today. This approach led to the choice of Cisco 7600 for the edge networks and Cisco 12000 for the core allowing the de facto construction of a 2.5 Gbit/s QoS capable network. Such a choice for the routers has to be complemented by the choice of the state of the art QoS Line cards both for the 7600 and the 12000 in the MB-NG testbed.

The Cisco 7600 is more flexible than the 12000 and thus employed at the edges. The 7600 presents the ‘ Parallel Express Forwarding (PXF) IP Service Processor’ which ensures the implementation of QoS functionalities such as scheduling and traffic shaping. It does this directly in hardware using an array of 16 programmable processors for parallel processing [Cisco1], [Cisco2], [Cisco3].

The Cisco 12000 - less flexible than the 7600 but much more powerful and thus employed in the core - implements a leading edge internal architecture which ensures the absence of Head of line blocking (HOL) by the use of virtual output queues (VOQ). An input and output blocking dramatically reduced by the use of VOQ- prioritization and speedup are also provided [Cisco4].

Such technology finally enables the idea of deployment and research on IP QoS at 2.5 Gbit/s and over to be realized.

The rationale behind all the tests we conducted is based upon:

- Bottleneck imposition
- Per-class metric definition and measurement

The presence of a bottleneck ensures congestion which in turn can trigger the bandwidth-guarantee scheduler whose performance is of critical importance in any IP QoS deployment process. Such bottleneck can be physical (direct) or virtual (indirect) as discussed in Section 4.3. The metrics to measure are to be defined on a per class basis since each class is built on the QoS requirements of specific types of application traffic. For example, a class associated with an application which only attempts to utilise the unused bandwidth in a non-invasive a manner as possible, certainly presents totally different requirements than those of say a videoconference. The generic set of possible metrics is:

- Throughput
- Latency
- Jitter
- Loss
- Reordering
- RTT

Two broad areas in the deployment of QoS in IP networks have been envisaged:

One studies network dynamics and relative QoS solutions when the bottleneck(s) are in the core while the other makes instead the assumption that the bottleneck(s) are present in the edge or boundary networks which surround an over provisioned core.

The end-to-end QoS that two users receive is nothing but the concatenation of the QoS treatment received in all the routers along the path connecting them. Therefore, an end to end QoS deployment study has to firstly focus on the standalone performance of all the router types employed in the network prior to the actual engineering phase of building an end-to-end QoS service across multiple domains.

4.2 Standalone tests on the Cisco12000 series router

An extensive series of standalone tests have been conducted on the Cisco 12000 core router platform. A standalone test of an IP router, as per definition, implies that no layers above the IP-layer are used if the “Heisenberg effect” of the test equipment itself affecting the observation is to be avoided. In order to meet such requirement, a dedicated hardware test equipment from Spirent (the Adtech AX4000) is used which also gives us the great advantage of reaching the maximum theoretical line rate for any packet sizes.

For the detailed report of the tests on the Cisco 12000 router, please see [DOC1]. The aim of these tests is to provide the standalone performance measurements of the router in such a way that when they are connected together in the MB-NG core network, the composite behaviour can be understood. We are interested in the maximum packet forwarding rate, the maximum throughput and the latency, all as a function of the packet size. An indication of the queue lengths is also useful.

The test layout is shown in Figure 4. This shows an Adtech AX4000 connected to a Cisco 12000 series router. The traffic is transmitted from the AX4000 through the router and received back in the AX4000 and analysed. Two types of interface cards are used as the input and output of the 12000. They are Engine 2 and Engine 3 interface cards.

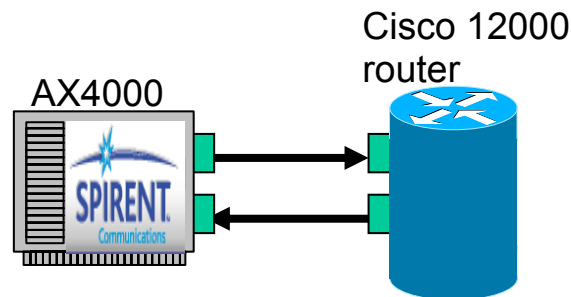


Figure 4. The setup for the standalone tests.

The results of the tests are shown in Figure 5 and Figure 6. Figure 5 shows maximum throughput for Engine 2 to Engine 3, Engine 3 to Engine 2 and Engine 2 to Engine 2+

(The Engine 2+ has 256Mbytes compared to 128 for the Engine 2, but otherwise they were the same). For these tests, we found that the extra memory did not have any noticeable effects.

All card combinations achieved line rate above 83 byte packet size. However, we note that the Engine 3 to Engine 2 achieves the highest packet rate – 3.84 million packets/s compared to 3.54 million packet/s for all other combinations.

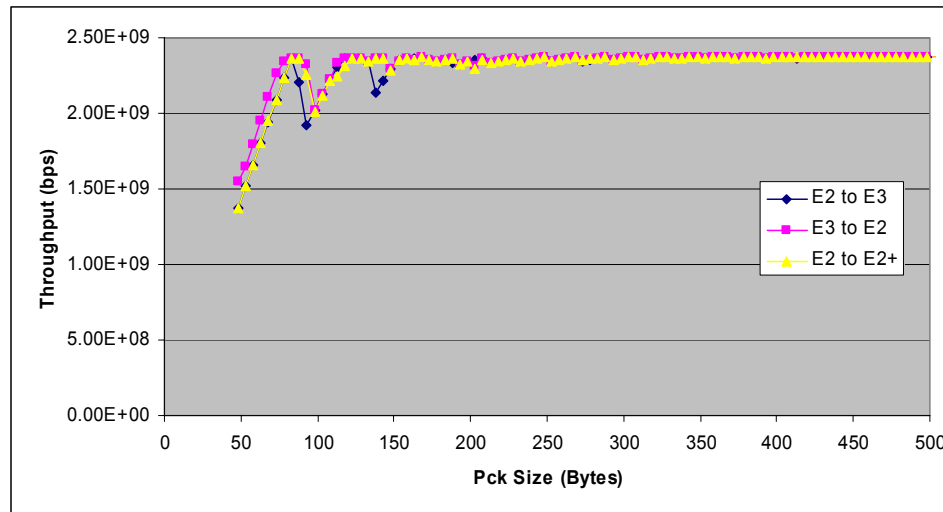


Figure 5.The maximum throughput for the Cisco 12000 for various combinations of line cards.

Figure 6 shows the packet size against the router port-to-port latency. We see that the Engine 2 to Engine 2+ has the smallest latencies. The Engine 3 to Engine 2 has an extra four microseconds and the Engine 2 to Engine 3 has an extra seven microseconds.

The peaks in the latencies which correspond to the troughs in the throughput are known as cell dips. These are caused when the packet size is just larger than the router's internal MTU, leading to less efficient transmission across the router's switching fabric.

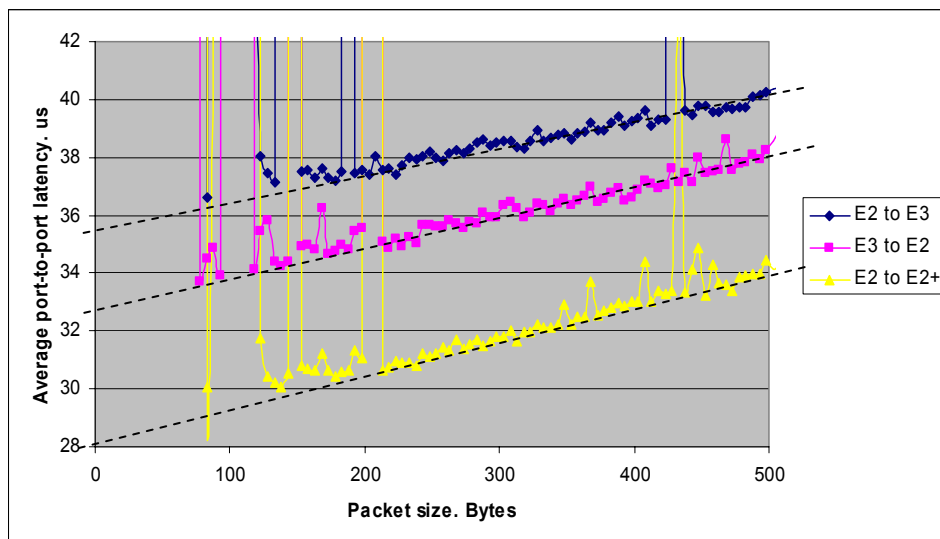


Figure 6. The port-to-port latency through for the Cisco 12000 for various combinations of line cards.

These results are summarised in Table 1. The dotted lines shown in Figure 6 take the form:

$$Latency(\mu s) = Packet_Size(Bytes) * m + c$$

The variables m and c are given in Table 1 for each combination of line card.

	Peak throughput. Gbit/s	Peak packet rate. Millions of packets/s	Minimum packet size for peak throughput/package rate. Bytes	Latency variables.	
				$c(u s)$	$M(us/byte)$
E2 & E2+	2.36	3.54	83	28.5	1/89
E2 to E3	2.36	3.54	83	35.5	1/117
E3 to E2	2.36	3.82	83	33.0	1/98

Table 1. Summary of the Cisco 12000 standalone performance

4.3 End-to-end tests: CORE bottleneck test results

As already mentioned in the introduction, one of the possible scenarios whose dynamic has to be studied and understood for the deployment of QoS in the MB-NG network is a bottleneck in the core-network, that is, in the Cisco 12000 routers at the core of the MB-NG network. The core bottleneck tests results that follow consist of two parts:

- A virtual core bottleneck – Using shaping to emulate a reduction of the speed of the interface. Also known as spacing.
- A physical core bottleneck – Using an OC-3 interface running at 155 Mbit/s.

4.3.1 Virtual bottleneck: QoS_On_Shape

For a detailed report of these tests please see [DOC2]. The aim of these tests is to investigate QoS as a solution to differentiate traffic traversing a bottleneck in the core (nominally in a Cisco 12000) for a broad choice of different-sized bottlenecks.

The constraint which makes this difficult in a real testbed project like MB-NG is that it is not practical and cost effective to buy n cards to have n different-sized bottlenecks especially when this choice would not give us the sizes-granularity we would like to have.

The solution is to emulate a bottleneck by making one of the 2.5 Gbit/s interfaces of the Cisco 12000 act as if it was less than 2.5 Gbit/s, in this way creating a bottleneck since the traffic would flow from a 2.5 Gbit/s interface to one of a lesser capacity. This maintains the real-testbed aspect which is the strength of the MB-NG project while at the same time offering the same flexibility in the test granularity we would obtain with software tools like ns-simulator.

We must stress that as well as being a technique to emulate bottlenecks of different sizes, this bottleneck-emulation technique is of great practical benefit in at least two real scenarios:

1. If an interface in the core is detected as a potential bottleneck of size k but QoS solutions are not available or do not work on such interface, then a bottleneck of size k can be emulated in one of the upstream interfaces where QoS works. This can be thought of as a sort of “pre-emptive” QoS deployment.
2. When the traffic traversing a router output port feeding a core network has to be shaped to a certain target rate in order to gain known statistical properties. In this way allowing network dimensioning and admission control to be done based on known Markov models [PAPER1]. We might want to treat all the classes sharing that output port shaped flow differently.

We refer to the bottleneck-emulation technique as the “QoS_On_Shape” method. The end-to-end network layout for the bottleneck-emulation QoS_On_Shape test is shown in Figure 7. This shows the UCL site streaming traffic to the Manchester site. The egress port of the Cisco 12016 GSR is used to emulate the bottleneck. There were three flows: one Best Effort (BE) flow and two Less than Best Effort (LBE) flows. Each flow run at 735 Mbit/s. The size of the bottleneck ranged from 700 Mbit/s to 2.5 Gbit/s.

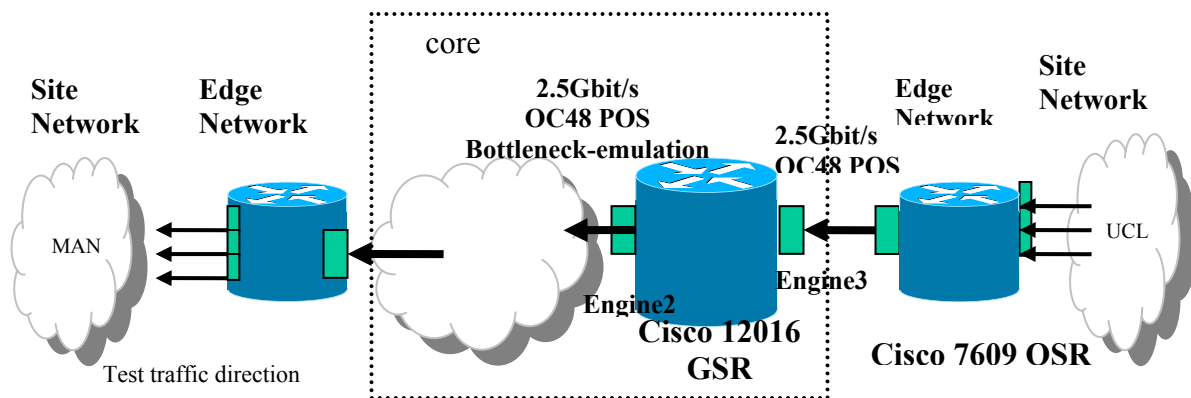


Figure 7. The setup for bottleneck emulation tests

In engineering the QoS to work on the specified interface, we enabled Modified-Deficit-Round-Robin (MDRR) to perform the packet scheduling and hence the bandwidth allocation. MDRR required Weighted-Random-Early-Detection (WRED) enabled to perform packet dropping.

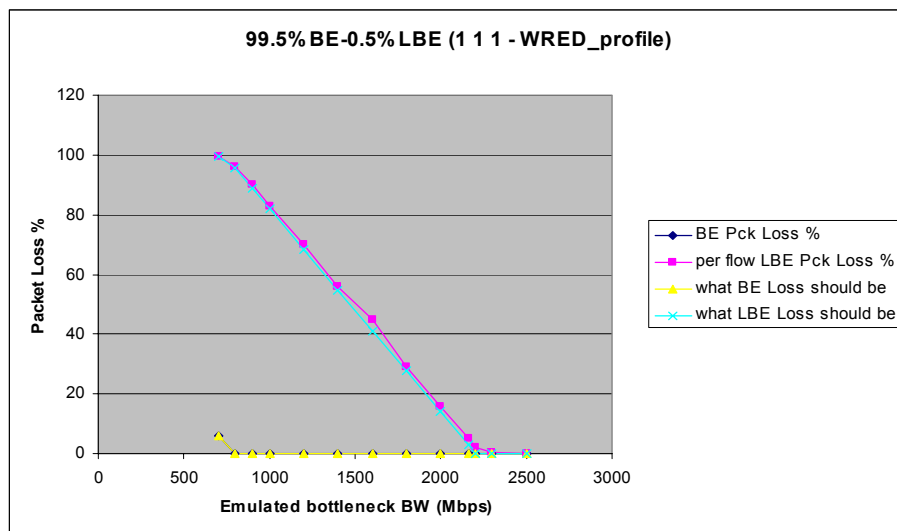


Figure 8. The packet loss for BE and LBE classes for various bottlenecks.

In Figure 8, 99.5% of the bandwidth reserved for BE and 0.5% reserved for LBE at congestion. The figure shows the bottleneck size against the packet loss for each class. The results agree very well with that predicted. The WRED drop probability curves were the same for both classes and set to the most aggressive. The WRED minimum threshold was set to one and maximum threshold was also set to one with a drop probability of 100%. In what follows, we investigate the WRED settings.

In Figure 9, the bottleneck was fixed to 700 Mbit/s with 99.5% of the bandwidth reserved for best effort (BE) and 0.5% reserved for less-than best effort (LBE) at congestion. The WRED drop probability curves were set the same for both classes. The minimum threshold was fixed to one and the maximum threshold was made variable from 1 to 300,000 (the x-axis of Figure 9) with a drop probability of 100%. The effects on the two classes were looked at from the packet loss point of view. The expected results are compared with the experimental results. This shows the region where the MDRR settings are effective is with the WRED profile's maximum threshold below 50,000.

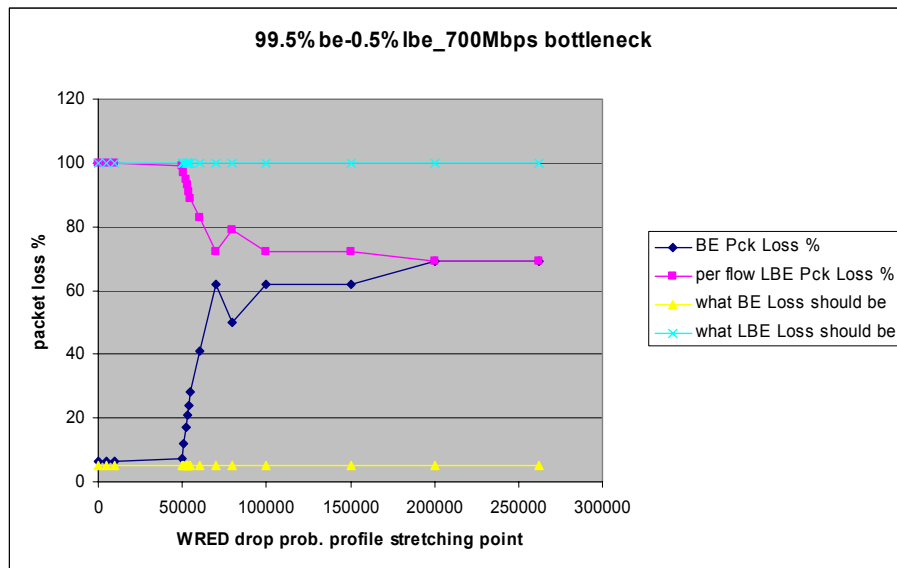


Figure 9. The packet loss for BE and LBE classes with different WRED profiles.

When the bandwidth was equally allocated to the two classes (using MDRR's maximum settings of 2048 for BE and 2048 for LBE) and the experiment repeated, we observe a different plot shown in Figure 10. The figure shows that the MDRR settings are effective for the WRED profile's maximum threshold setting in the region between 10 and 80,000.

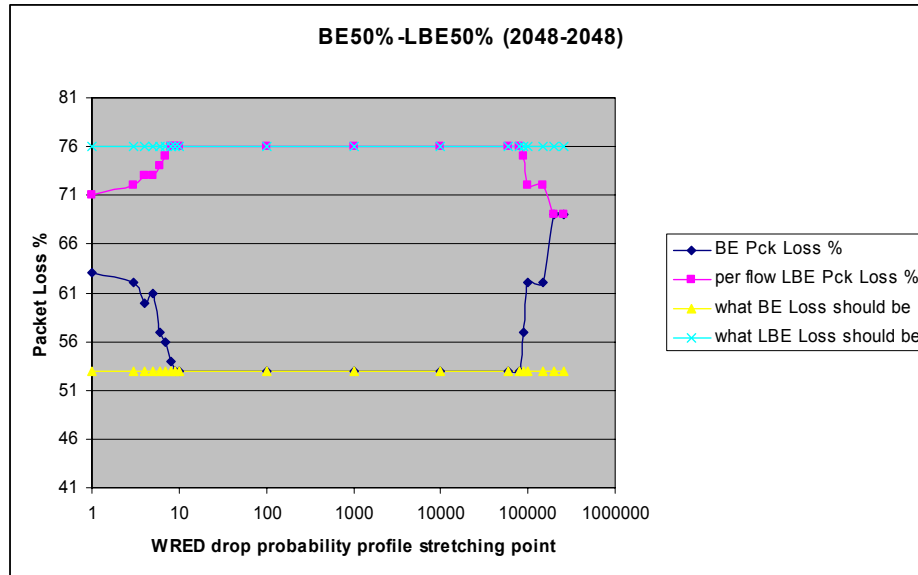


Figure 10. The packet loss as a function of WRED profiles for an MDRR setting of 2048 BE, 2048 LBE, i.e. 50% BE and 50% LBE.

The tests performed successfully demonstrated that the bottleneck emulation in the core is possible and that differentiating the traffic traversing it through the use of MDRR and WRED (QoS_On _Shape) is also possible on condition that WRED is properly engineered.

We found three operating regions:

1. Where the WRED drop probability profile's maximum threshold is below 10. In this region, WRED is very aggressive in dropping packets and the differentiation does not work for high values of MDRR (above 1000) in equal (50%-50%) bandwidth allocations. However, it works well for any values of the quanta when the BW allocation is not equal (99%-1%, 98%-2% etc).
2. Where the WRED drop probability profile's maximum threshold is above 50,000. In this region, WRED is not aggressive enough in dropping packets and the differentiation does not work.
3. Where the WRED drop probability profile's maximum threshold is between 10 and 50,000. This is the region where the WRED drop probability profile enables the MDRR settings to be effective.

It is worth highlighting that the performance of the system when congested are very good in both the following critical test situations:

- When the BE class load over-subscribes the BE minimum BW guaranteed, both classes show that the two minimum BW guarantees are met.

- When the BE class load under-subscribes the BE minimum BW guaranteed, LBE is able to get the whole remaining unused BE BW, in this way allowing the system to always perform at the maximum link utilisation, which reveals the ability of the MDRR algorithm in reallocating the unused BW.

4.3.2 Physical bottleneck: OC3

For a detailed report on these tests please see [DOC3]. The main aim of this tests is to investigate QoS differentiated traffic traversing a real bottleneck in the MBNG core, nominally in a Cisco 12000.

We placed a POS OC-3 Line Card, whose line rate capacity is of 155Mbps, between the MB-NG core and UCL edge network as shown in Figure 11. Two flows - a BE and an LBE flow - were injected from Manchester to UCL at different loads, from 100Mbps up to 2.0 Gbit/s

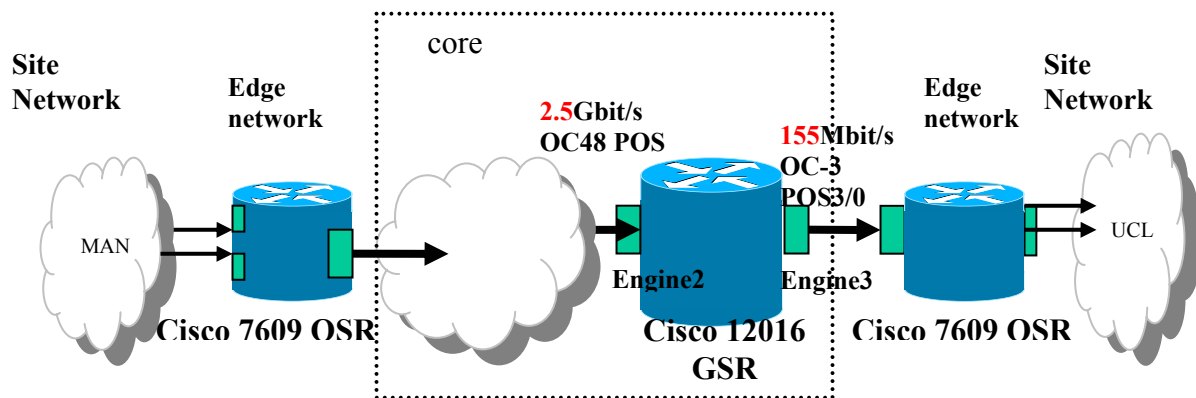


Figure 11. The setup for testing the network with a physical bottleneck

Policies allowing the following bandwidth allocations between BE and LBE were implemented on the OC-3 interface of the Cisco 12016 core router: 98%-1%, see Figure 12 for results; 69%-30%, see Figure 13 for results; 53%-45%, see Figure 14 for result. The figures show the offered load against the achieved throughput. Below the congestion point, each class gets an equal share of the bandwidth. Above the congestion point, BE is favoured over LBE at exactly the configured bandwidth allocation.

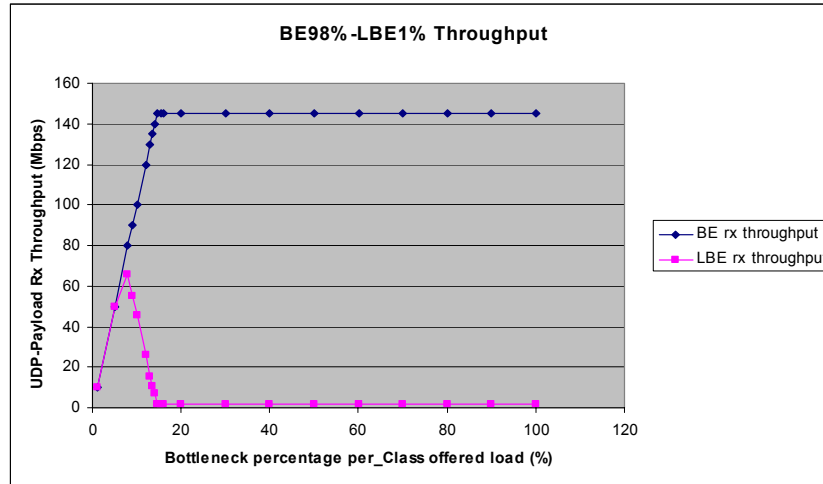


Figure 12. Offered load against achieved throughput for bandwidth allocation of 98% BE and 1% LBE.

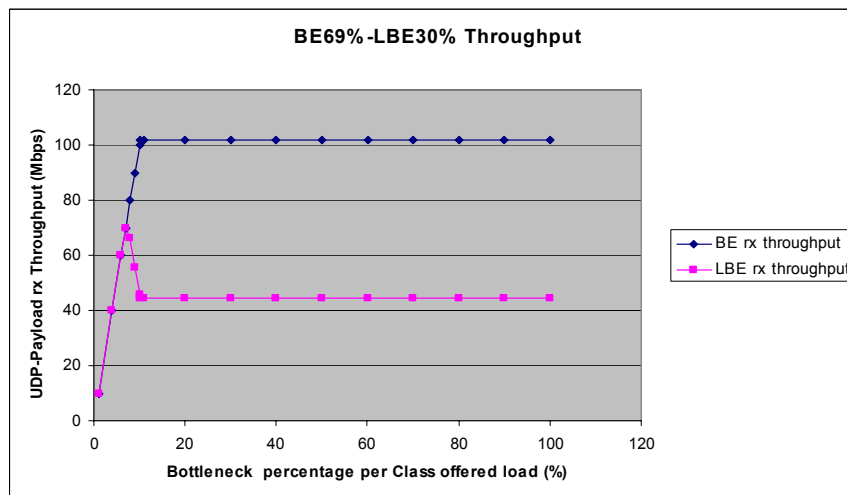


Figure 13. Offered load against achieved throughput for bandwidth allocation of 69% BE and 30% LBE.

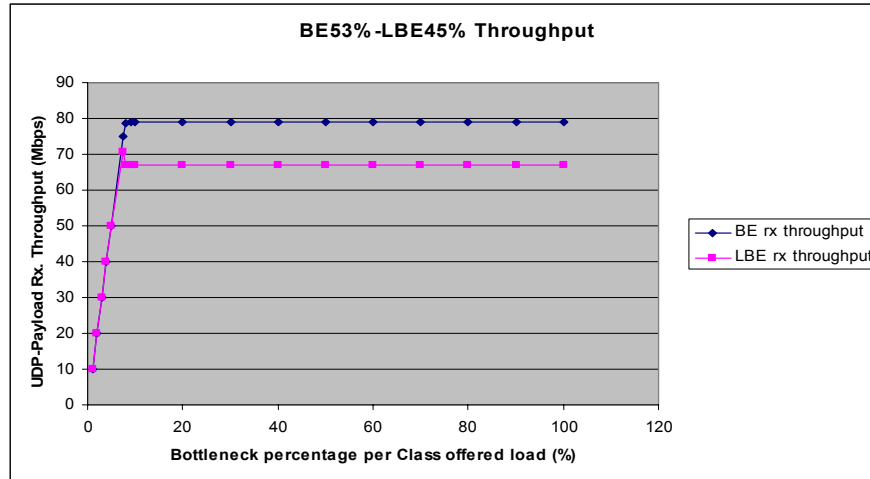


Figure 14. Offered load against achieved throughput for bandwidth allocation of 53% BE and 45% LBE.

In these tests, we successfully engineered the Quality of Service techniques in the extreme ($155/2000 = 1290\%$ oversubscribed) situation where a physical bottleneck interface of 155Mbps capacity is placed in a Cisco 12000 2.5Gbps core IP-network. We proved that a successful QoS implementation is definitely possible if the mandatory joint usage of the Modified-Deficit-Round-Robin (MDRR) and Weighted-Random-Early-Detection (WRED) algorithms is well engineered. Based on our investigation of the WRED parameters in the Section 4.3.1, the following WRED parameters were used: minimum threshold of 1 and a maximum threshold of 1 with a drop probability of 100%. The observed results clearly validate our finding in Section 4.3.1.

The parameter space of the BW allocation settings of 98%-1%, 69%-30% and 53%-45%, showed very good performances in both the following critical test situations:

- When BE class load is under-subscribed, LBE is able to get the whole remaining unused BE bandwidth. This reveals the MDRR algorithm's ability to allow the system to always perform at the maximum link utilisation by reallocating unused bandwidth.
- When link is over-subscribed, both classes show that the minimum bandwidth guarantees are met.

The tests also proved the absence of any explicit (direct) or implicit (indirect) form of backpressure when at least one of the two interfaces involved is configured with a Quality of Service policy [DOC3].

4.4 End-to-end tests: EDGE bottleneck

As already mentioned in the introduction, one of the possible scenarios whose dynamic has to be studied and understood for the deployment of QoS in the MB-NG network is that with the bottleneck in the edge networks. In MB-NG terms, this is the scenario

where the congestion point is in a Cisco 7609 OSR Router. The setup used to test this scenario is shown in Figure 15. Test traffic flowed from UCL to Manchester. There were two BE flows and one LBE flow. The QoS policy was applied on the egress of the Cisco 7609 OSR on the UCL side. For the 7609, we were required to use weighted fair queuing (WFQ) in order to implement the policies.

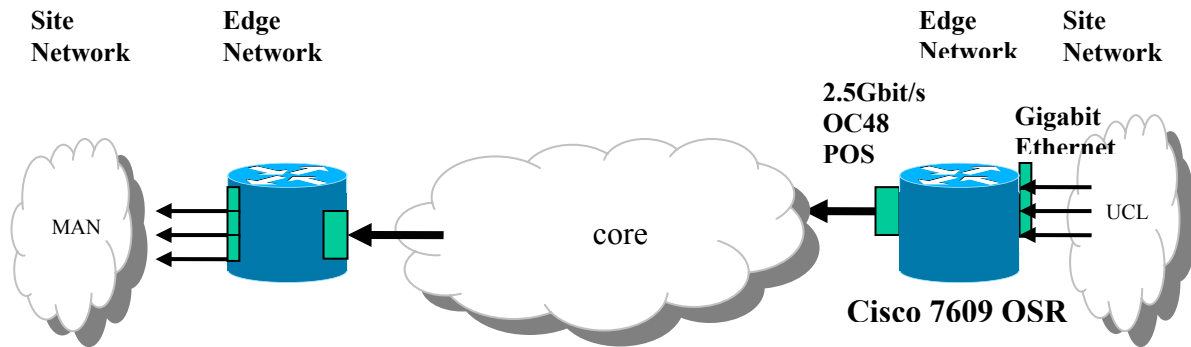


Figure 15. The setup for testing QoS when the congestion is at the edge network.

Figure 16 shows the offered load against the achieved throughput for the bandwidth allocation of 93% for BE and 6% for LBE. Below the congestion point, each class gets an equal share of the bandwidth. Above the congestion point, the BE flows is favoured over LBE flow.

The system performs as expected. As in the case for the congestion in the core, when the BE class is under-subscribes the LBE class is able to get the unused BE bandwidth allowing the system to always perform at the maximum link utilisation percentage which reveals a well behaved WFQ algorithm.

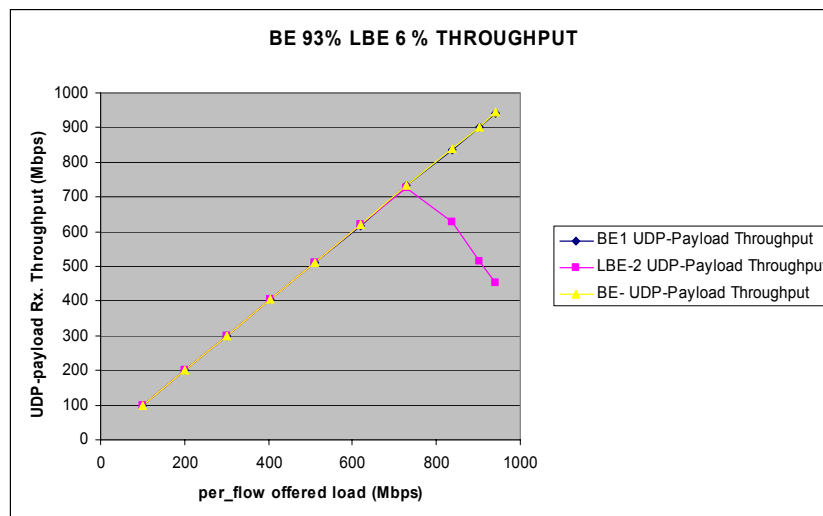


Figure 16. Edge network bottleneck test. Offered load against achieved throughput for bandwidth allocation of 93% BE and 6% LBE.

4.5 QoS - summary

The first engineering stage before the planned extension to a multi-domain environment and hence towards the final deploying of an end-to-end IP QoS service across multiple domains has been reached. By the use of constant bit rate traffic for different loads and different packet sizes, all the main components of the MB-NG edge and core networks have been engineered for QoS through the analysis of throughput and loss both in the standalone mode and in an end-to-end connected mode.

A first demonstration of an end-to-end QoS service for traffic traversing multiple domains has been demonstrated by deploying the state of the art QoS-capable Cisco routers at 2.5 Gbit/s.

Leading QoS techniques such as traffic shaping, MDRR, WRED and WFQ presented here have been engineered.

4.6 Future work

Delay-sensitive classes will be configured and engineered on the basis of accurate end-to-end measurement of metrics such as latency and jitter.

With the imminent extension to a multi-domain environment, close liaison with the MB-NG middleware group will be ensured in order to produce the SLA specifications between MB-NG network domains and to provide the admission control mechanism necessary to enforce such SLAs.

The end-to-end QoS service will be proved for real application traffic competing with all different sort of background traffic which can be real or artificial. MPLS will be deployed in the core in order to empower its traffic engineering capability as well as its flexibility in providing QoS.

References

[DOC1] http://www.mb-ng.net/private/year1_report/MB-NG-Doc-T8-0.1-GSRCharacterisation

[DOC2] http://www.mb-ng.net/private/year1_report/MB-NG-Doc-T8-0.1-Core_«QoS_On_Shaping»

[DOC3] http://www.mb-ng.net/private/year1_report/MB-NG-Doc-T8-0.1-OC-3-Core_Bottleneck

[PAPER1] http://www.tu-harburg.de/et6/papers/papers_desc_abendroth02c.html

[Cisco1] <http://www.share.org/proceedings/sh98/data/S3809.PDF>

[Cisco2]

http://www.cisco.com/en/US/products/hw/routers/ps368/products_white_paper09186a00800886d7.shtml

[Cisco3] http://www.cisco.com/warp/public/cc/pd/rt/10000/prodlit/pxfw_wp.pdf

[Cisco4] <http://www.bcr.com/bcramag/1997/12/mckeown.asp#4.5>

5 Middleware

In this section we report on the first steps made towards identifying and deploying middleware capable of giving access to network resources from the application level. This part of the project is at an early stage. Its completion will form one of the main objectives of the second year.

5.1 Overview.

Middleware between the Network layer and the Application layer is needed to allow applications to access the network resources. At the Network layer, it is assumed we already have a network supporting IP “Quality of Service” (QoS) mechanisms like “Differentiated Services” (DiffServ).

A middleware solution will provide an API to the applications, allowing them to access the network resources. An example of accessing network resources is to configure the edge router to accept a flow at a given data rate and marking packets in a specific Diffserv Code Point (DSCP). One or more policies are applied to the marked packets end-to-end across different domains. The API is required to support both immediate and advance reservations of the network resources.

5.2 Middleware Survey

Different middleware projects have been reviewed:

- **QDM & QPM:** “QoS Device Manager” and “QoS Policy Manager” are both Cisco products and interesting for management of the policies. However, it is not easy to implement API functionality for the application level because they are Cisco proprietary products and we do not have access to the code.
- **AQUILA:** This is a project in the EU IS framework and is in process of being developed and not compatible with the MB-NG deadline.

- **GARA:** The General-purpose Architecture for Reservation and Allocation is an open source project with version 1.0.4 already available. The current stable version of GARA requires Globus version 1.1.3. It fits in well with the aim of MB-NG for accessing network resources from applications.
- **UvA AAA Server implementation:** The University of Amsterdam's implementation of the AAA server. The initial version of the code is available. It fits in well with the aim of MB-NG, especially in the "Managed Bandwidth" task.
- **GRS:** "Grid Resources Scheduling", develop by Computer Science at University College London (UCL). The project is at the initial stages and may not be compatible with the MB-NG deadline.
- Other projects: "**Open ORB**" from Lancaster University, "**KUBB**" the Kansas University Bandwidth Broker, or the "**Bandwidth Broker**" from Merit Networks.

For initial studies we selected the GARA and UvA AAA Server implementations for deployment upon the MB-NG testbed.

On the 20th and 21st February the MB-NG project organized a middleware workshop at UCL. A wide audience, including people outside the project attended. The principal authors and project leaders for GARA (Volker Sander) and UvA AAA (Leon Gommans and Bas van Oudenaarde) were able to attend to give in-depth presentations of their work and answer questions. Different options and models were reviewed for both products. For each model we identified what was missing in each product and what needed to be developed.

It was widely acknowledged that the two products can complement each other and be integrated into an environment such as MB-NG. We can say that GARA supports immediate and advance reservations, provisioning and configuring edge devices to allocate resources. However GARA is missing the accountancy and authentication and at the moment relies on Globus, therefore and integration of GARA with AAA will fill these drawback. Also AAA could deal better with the multi-domain scenario present in MB-NG.

5.3 GARA Status

The current stable GARA version is 1.0.4 and it must be built on Globus version 1.1.3. This version of Globus, though very stable and widely used, it is not the latest version of this toolkit. The Globus v3.0 has been launched recently, but the latest stable version of Globus is 2.2. GARA developers have just ported the GARA to the Globus version 2.2. Many features were improved so GARA is being updated to version 1.2.2.

5.4 Current Testbed

We have GARA version 1.2.2 with the Globus Toolkit version 2.2.4. The testbed we are using at the moment is quite simple but effective enough to verify that the GARA functionalities. The testbed is shown in Figure 18.

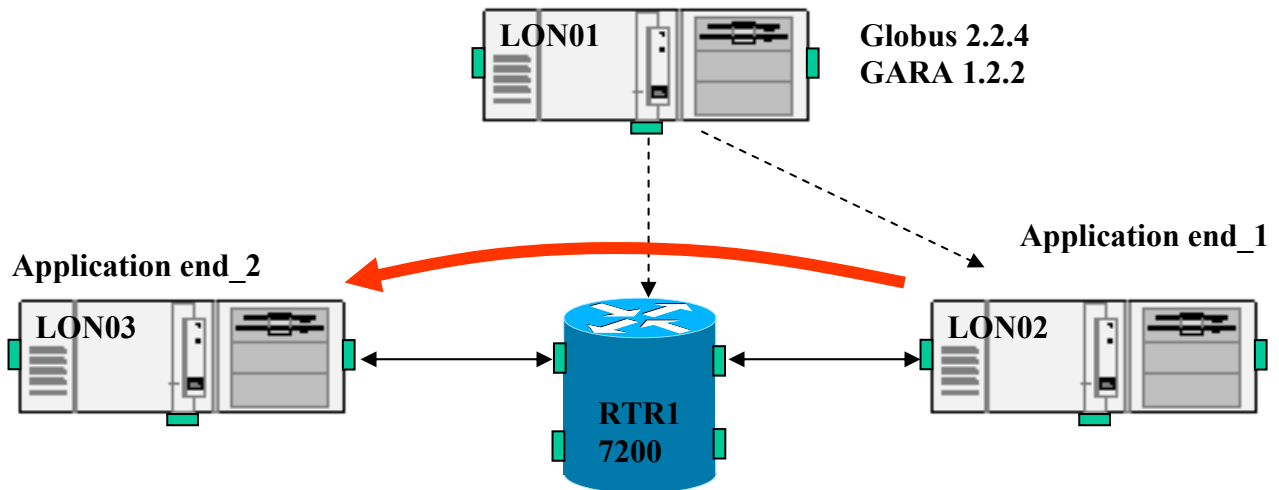


Figure 18. The testbed to verify the GARA functionality.
Figure 17: Basic testbed for Globus and GARA.

In Figure 18, a Cisco 7200 router is used to make the reservation of resources and connect LON02 with LON03. LON01 is used to run the GARA “Bandwidth Broker,” configuring the router and enabling the application layer (running at LON02 and LON03) to see network resources.

5.5 Future Activities

The next step will be to provide the same functionality but replacing the Cisco 7200 router by the entire MB-NG network and running end to end reservations between Manchester and London for instance, but still having a single domain and configuring the routers at the edges.

At that time we expect to have the integration between GARA and AAA and it would be possible to use the AAA functionality to do the accountancy and authentication. In this new context we can have multiple independent domains with related service level agreements (SLAs) and the AAA implementation should deal providing the end to end QoS and leaving to GARA the tasks of making resources available to the applications and configuring network devices as requested by the applications.

6 High throughput program

6.1 High Throughput Data Transfers

The initial aims of the High Performance project are:

- Determine the conditions needed to allow high throughput data transfers.
- Investigate the behaviour of various TCP stacks on high bandwidth links.
- Demonstrate high throughput memory-memory data transfers over long periods of time.
- Demonstrate high throughput disk-to-disk data transfers over long periods using a variety of data moving applications and transport mechanisms.

A schematic diagram of the network and end systems used in the tests is shown in Figure 19.

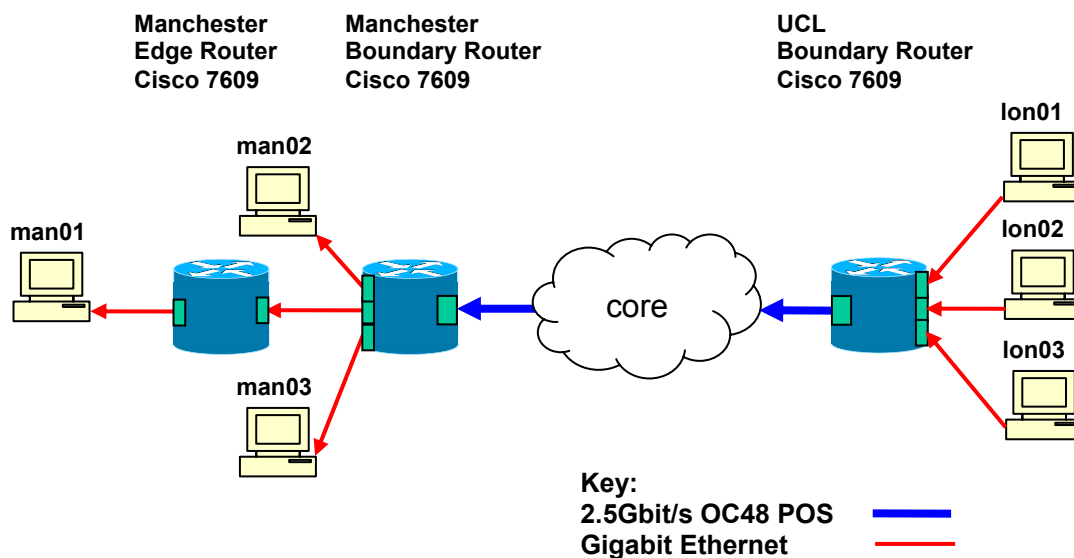


Figure 19. The routers and end hosts used for the high throughput tests.

6.2 What's Involved in Achieving High Performance Networking

There are many factors that determine end-to-end network performance, these include:

- The Applications: The architecture and the implementation can determine how an application accesses and uses the network and i/o subsystems. For example, the use of blocking calls may be appropriate in one case and introduce considerable delays in others.
- The End Hosts: The choice of CPU power, motherboard (here it is the chipset and PCI bus that are most important) and NIC all influence the measured performance and packet dynamics.

- The Routers: The architecture, the facilities available, the configuration selected and the performance of the blades and switching fabric all impact the end-to-end behaviour.
- The Network itself: Clearly the Bandwidth of the links, the amount of traffic and the presence of components like firewalls impact on the throughput observed, but the framing or way the data is moved is also important as this introduces overheads.

6.3 Some Issues Affecting Performance

Initial throughput measurements using UDP gave very variable throughput and suffered from considerable packet loss. A typical plot of throughput taken over a three hour period is shown in Figure 20. The TCP throughput for long periods was correspondingly poor and the throughput from London to Manchester was not equal to that from Manchester to London. This behaviour was not expected for the MB-NG development network operating with no bottlenecks and cross-traffic.

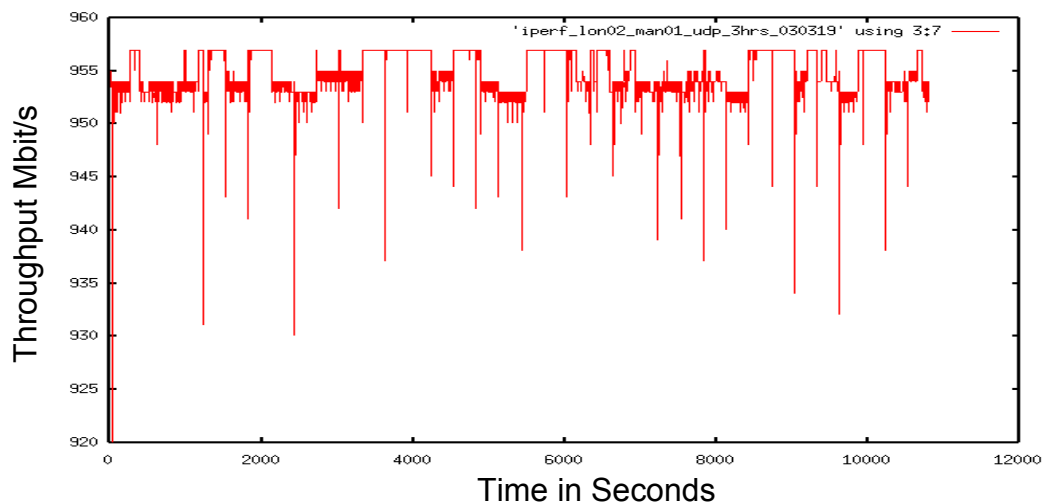


Figure 20. An initial time series plot of the throughput obtained using UDP over a three hour period. It shows very variable throughput.

6.3.1 Investigation of Packet Loss

By checking the packet counters in the kernel of both hosts, it was found that packets were being lost in the receiving host. They were being discarded when the application or higher levels of the TCP stack could not process the packets as fast as they arrived. When the NIC interrupted the CPU on the receipt of every packet, there are considerable variations in the throughput observed as shown in the top plot in Figure 21 which shows how the throughput varied with the interrupt coalescence. There was also significant packet loss within the receiving host when the NIC interrupted the CPU every packet, indicating that large amounts of CPU power were being used in servicing the interrupts.

With receive interrupt coalescence of 20 or 80 μ s the variation was reduced to ± 1 Mbit/s and no packets were lost.

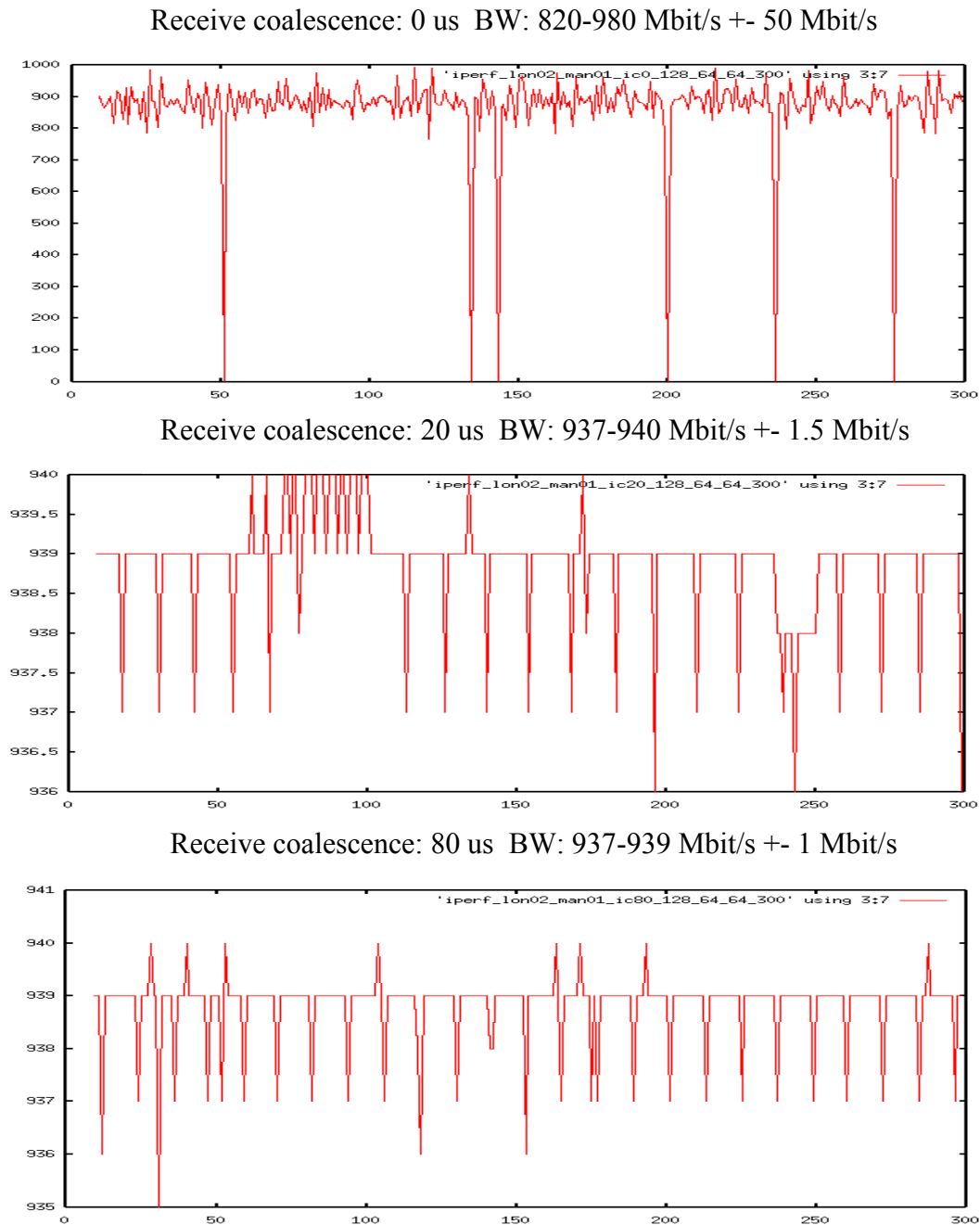


Figure 21. Throughput in Mbit/s vs time for different values of the receive interrupt coalescence.

6.3.2 Investigation of Errored Packets

It was also found that the NIC driver was reporting Errored packets. Errored packets were only reported in the receiving end host for packets coming over the POS link from London to Manchester, and not for packets send between hosts within Manchester. In

both cases the packets travelled over the same Gigabit Ethernet links in Manchester so the errored packets could not be due to Gigabit Ethernet framing errors on the local links. It was found that a pre-production OSM-1OC48-POS+SS WAN card was somehow changing the packets and this happened for packets less than 200 bytes. This unfortunately prevented reliable deliver of TCP ACK packets leading to very poor throughput! The rest of the data presented in this report used the older POS WAN card, not the pre-production POS+ version. Later tests showed that there were no error packets when the production POS+ WAN cards were used.

6.3.3 The importance of the Transmit Queue length

The transmit Queue is situated between the output of the IP stack and the driver for the individual NIC and contains data describing each packet that needs to be sent by the driver. It is the kernel that schedules the NIC driver to run and empty the queue, and this is independent of the TCP/IP stack. The length of the queue is determined by a kernel parameter called “txqueuelen”. If this queue becomes full, for example if the software can generate packets faster than they can physically be transmitted, the packets are not lost but some TCP stacks treat this condition as a *real* congestion signal and cut the sending rate. This condition is known as a “sendstall”. Figure 22 shows the number of sendstalls measured for different transmit queue lengths. To avoid sendstalls the txqueuelen was set to 2000 for the tests reported.

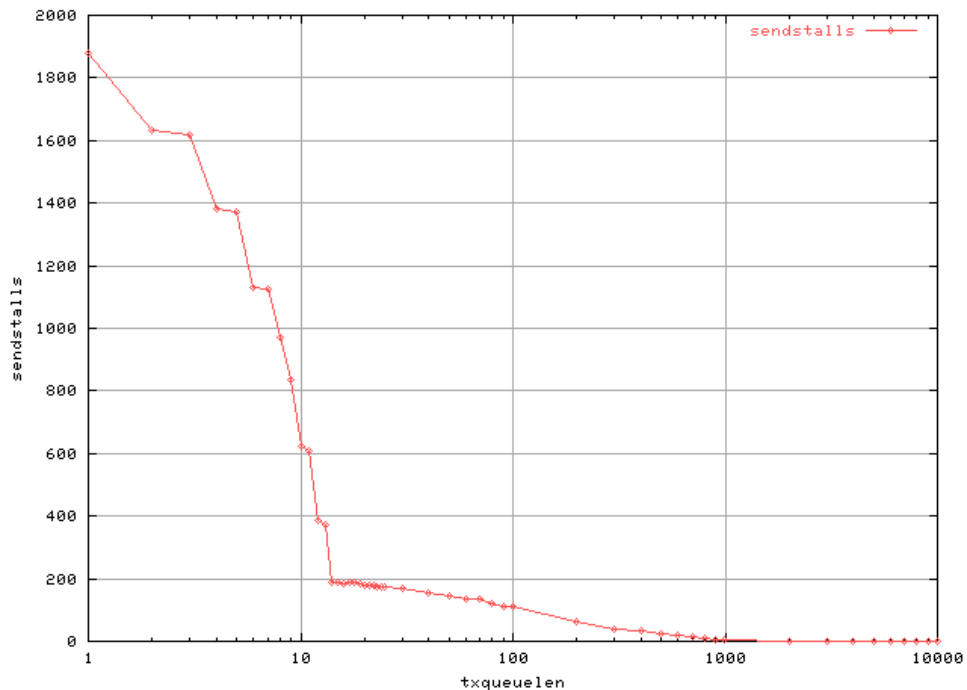


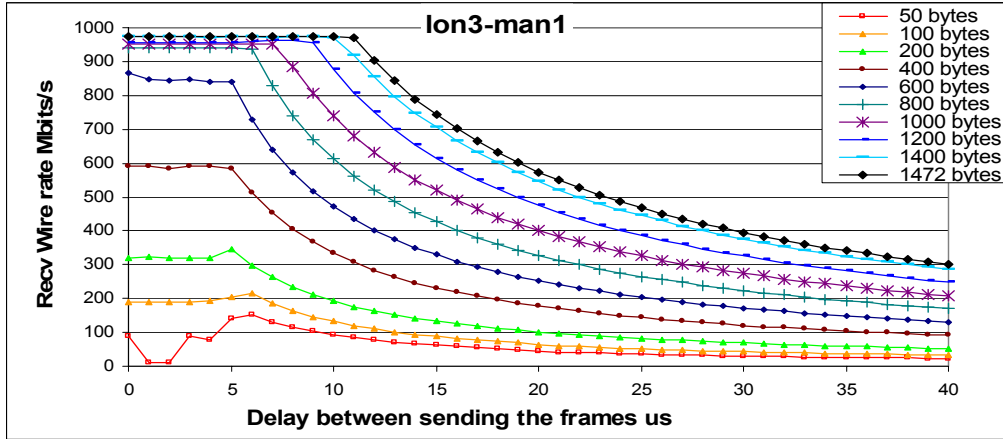
Figure 22. The measured variation in the number of “sendstalls” reported as a function of the transmit queue length “txqueuelen”.

6.4 End-to-End Memory-memory Tests with UDP

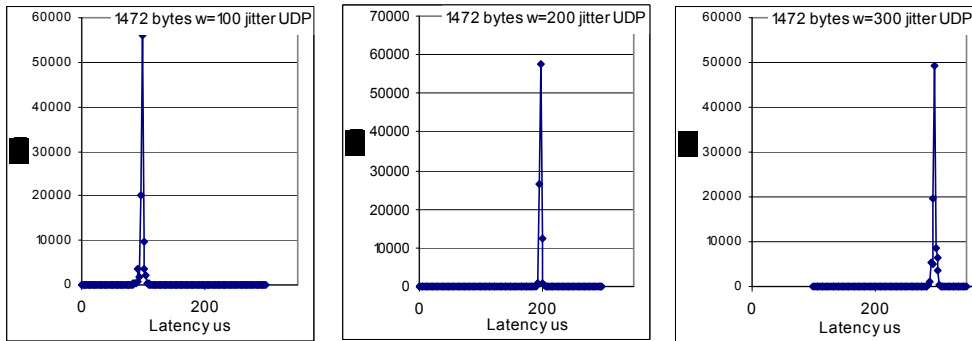
UDPmon [UDPmon] was used to check the basic performance of the network by sending streams of carefully spaced UDP packets. Figure 23 A shows the throughput as a function of the delay between sending the packets for various packet sizes. For packets larger than 600 bytes, the behaviour is as expected, with constant near line throughput (950 Mbit/s) and then a $1/t$ dependency when the delay between packets becomes greater than the time required to physically transmit the packet. For smaller packets there is some drop in throughput together with unexpected packet loss.

The histograms of packet jitter shown in Figure 23 B are clean with a Full Width Half-Maximum of $4\mu\text{s}$, indicating that there are no queuing effects in the network.

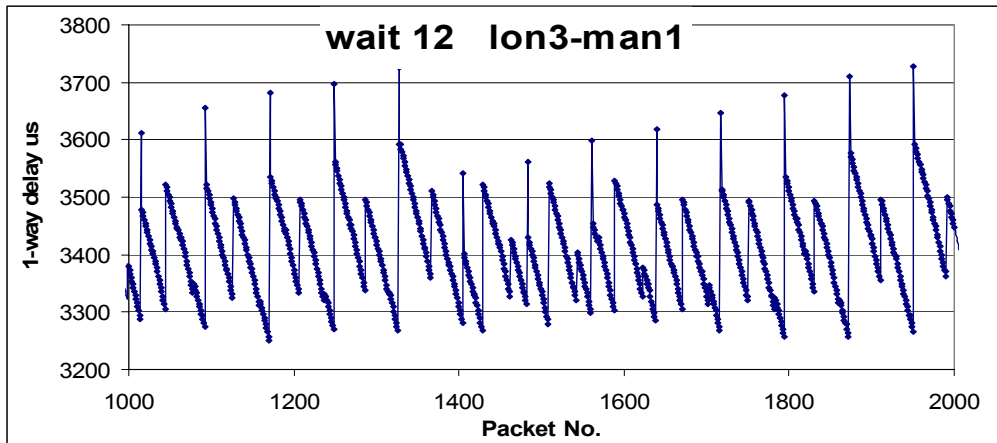
The saw-tooth behaviour of the 1-way delay estimates shown in Figure 23 C are typical of interfaces using interrupt coalescence where packets wait in the receiving host until the interrupt is given by the NIC some time later. Packets transmitted earlier wait for longer giving a longer transit delay. Again, there is no evidence of queuing in the network.



A. Throughput vs time between sending the packets



B. Packet Jitter 12 μ s between packets



C. 1-way delay as a function of elapsed time

Figure 23. End-to-end tests using UDPmon.

MB-NG has also examined the performance of new transport protocols that improve standard 'Vanilla' TCP. Figure 24 compares the throughput obtained from 'Vanilla' TCP, HighSpeed TCP [HSTCP] and Scalable TCP [ScalableTCP] when the network has different packet loss rates. The packet losses were induced artificially by modifying the

kernel on the receiver to periodically drop packets, thus simulating the effect of real network loss. Initial analysis shows that the two new stacks perform better under high loss rates than Vanilla or standard TCP. This is a first step approach to understanding the dynamics of TCP. Further investigations will be conducted under realistic network conditions in order to comprehend TCP behaviour and ways that this can be used to improve the performance of Grid applications.

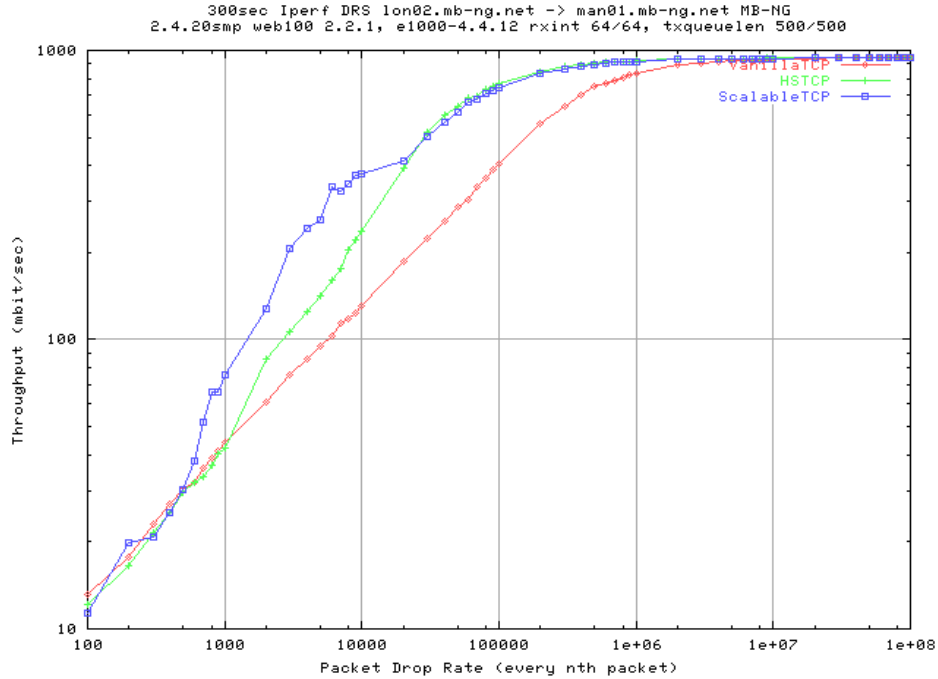


Figure 24. comparison of the throughput obtained from 'Vanilla' TCP, HighSpeed TCP and Scalable TCP as a function of the packet loss rate on the network.

6.5 End-to-End Memory-memory Tests with TCP

6.5.1 24 Hour Throughput with a Single TCP Stream

Using the HighSpeed stack, a single TCP stream was generated with iperf from lon02 to man01 and run for 24 hours. The time series throughput reported by iperf is plotted in Figure 25 with an expanded detail shown in Figure 26. They show an average user data throughput of 941.5 Mbit/s. The transmit interrupt coalescence parameters for the transmitting node, lon02, were set to 64 (the default for the Intel Gigabit Ethernet NIC) while the receiving node had Rx at 64 and Rx-absolute at 128 (the default being 0 and 128). The TCP buffer sizes were set to 1Mbyte (the rtt-BW product was 750 kbytes) and the txqueuelen was 2000.

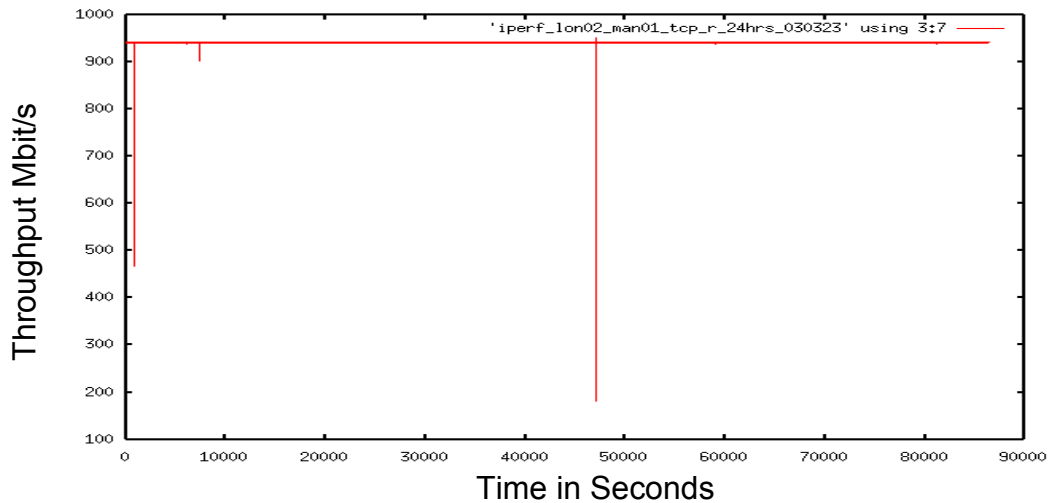


Figure 25. 24 Hour time series throughput for 1 stream using the HighSpeed TCP stack. The average throughput was 941.5 Mbit/s.

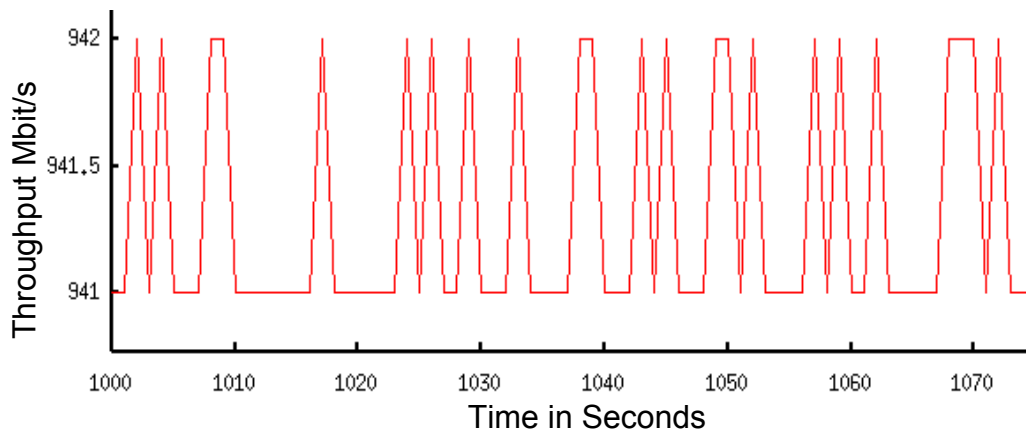


Figure 26. Detail from Figure 25 after 1000s of operation. The variation in throughput is due to the resolution of iperf not transmission problems.

The results show that sustained data flows at Gigabit rates are achievable, providing that “Server-quality” PCs are used and the correct selection of parameters is made. The parameters include: The transmit queue length, the interrupt coalescence values in the Ethernet driver, and the TCP transmit and receive buffer sizes.

6.5.2 Multiple TCP Streams to one Remote Host

Three TCP streams were generated with iperf between Manchester and London using the HighSpeed TCP stack. The first stream was started from man01 to lon02, 60s later a second stream was started from man02 to lon02, and 60s after that the third stream was started from man03 to lon02.

The throughput was measured using the information collected from web100 and is shown in the upper plot in Figure 27. The other plots show the throughput of the initial TCP

connection during consecutive 60s time slices when two and three TCP streams active. The average rates reported by iperf drop from 940 Mbit/s to ~500 Mbit/s with two streams to 300 Mbit/s with three streams. The 1 Gigabit Ethernet link into lon02 node is being congested and Figure 28 shows the TCP congestion window, the number of duplicate ACKs, and the number of SACKs received as the number of streams is increased.

When two Gigabit streams were sent the Gigabit link into lon02 becomes congested, and packets were dropped giving rise to the dramatic increase in the number of duplicate ACKs sent by lon2 to the transmitting node man1 – as seen in Figure 28 during the time from 6 to 12s. When three streams were sent, packets were lost on the 2.5Gbit POS link as well as the Gigabit Ethernet link into lon02.

These tests show that highspeed TCP does share the available bandwidth in an equitable manner when multiple streams are experiencing a constricted or a link with packet loss.

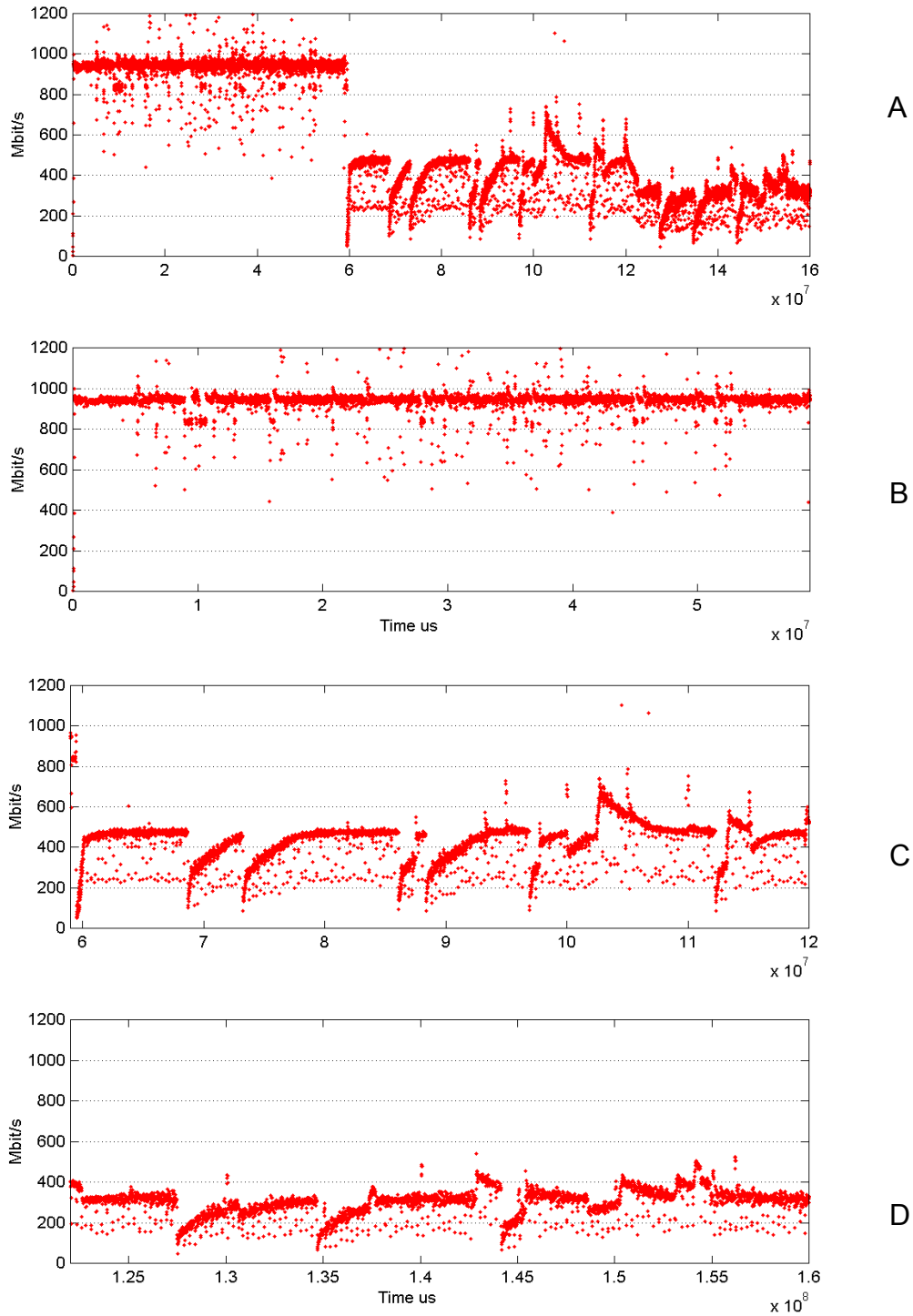


Figure 27. Time series throughput between man01 and lon02.

- A.** The full test period, at time $t=0$ one stream is started, 60s later a second stream is sent and then 60 later the third stream is started.
- B.** The throughput of the initial stream
- C.** The throughput of the initial stream when there are two flows
- D.** The throughput of the initial stream when 3 streams are sent to lon02.

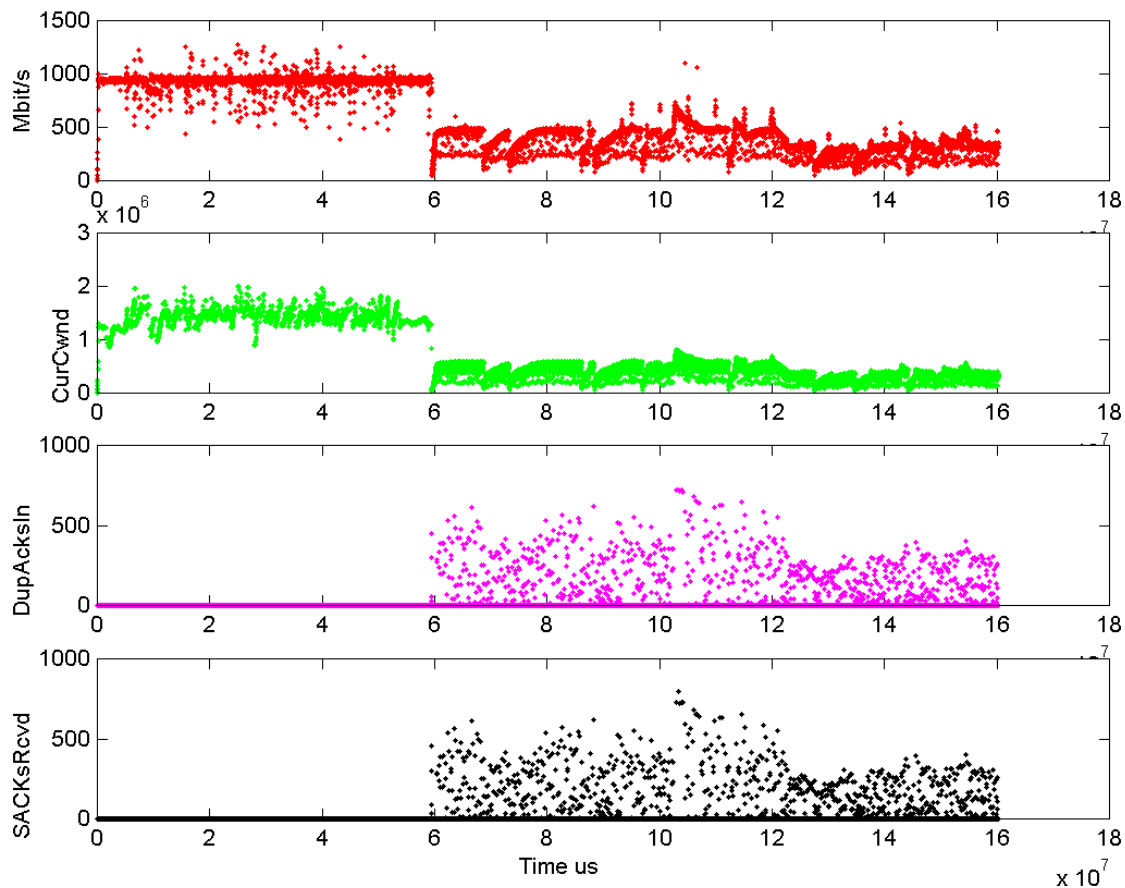


Figure 28. The TCP Congestion window, the number of Duplicate ACKs and the number of SACKs received corresponding to the TCP flows shown in Figure 27.

6.6 GridFTP Data Transfers

6.6.1 Throughput Measurements

Hardware Raid0 disk sub-systems were installed on lon03 and man01. The sub-systems used 4 Maxtor IDE disks, each connected with an ATA133 cable to the 3Ware 7500-8 controller which was a 64bit 33 MHz PCI card. The disk-memory throughput was measured using a 2 GByte file and gave 130 Mbytes/s for reading and 100 Mbytes/s for writing. This is consistent with the figures given by the manufacturer [3ware].

The data rates generated on the network by the data transfers were measured by recording the bytes sent and received by the network interface on the sending host every second. The rates obtained when multiple sequential transfers were made over a 30 minute period are shown in Figure 29. The lower plot (in green) shows the throughput for data leaving the host which was typically 520 Mbit/s; this rate includes the user data, as well as the TCP protocol and framing overheads. The upper plot (in red) shows the data received by the NIC and is due to the TCP protocol acknowledging the data that has been received by

the remote host. Similar results were obtained for PCs connected back-to-back indicating that the low throughput is not due to the development network.

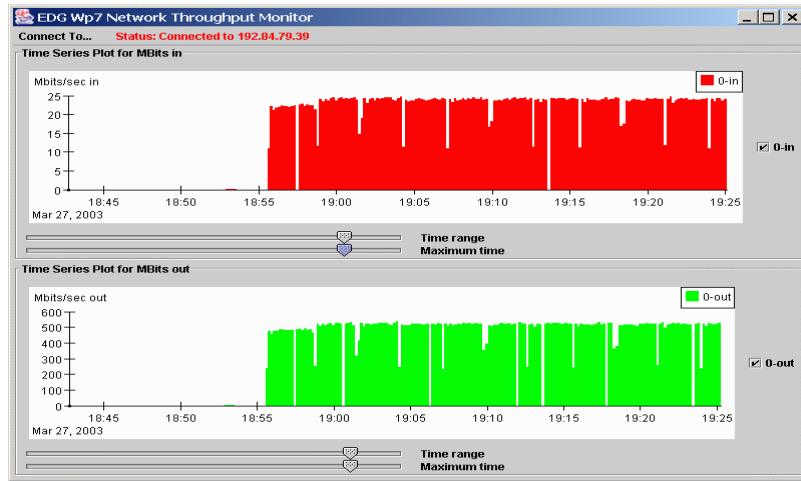


Figure 29. Time series of the network data rates for GridFTP transfers, these values were recorded from the Gigabit Ethernet interface.

6.6.2 Analysis of the Transfers

Figure 30 shows plots of the Web100 variables recorded every 10 ms during one GridFTP transfer. There are two bands of high throughput just over 800 Mbit/s and ~ 500 Mbit/s. The expanded data throughput plots show that the TCP transmit rates ranged from 500 / 800 Mbit/s to 1 Gbit/s but that alternate 10 ms samples showed zero throughput. This behaviour is not understood, but would account for the low throughput rates shown in Figure 29 which are the average over 1 second periods. This effect has not been seen in the other tests and is most likely due to the detailed behaviour of the GridFTP application. Discussions are in progress with the developers.

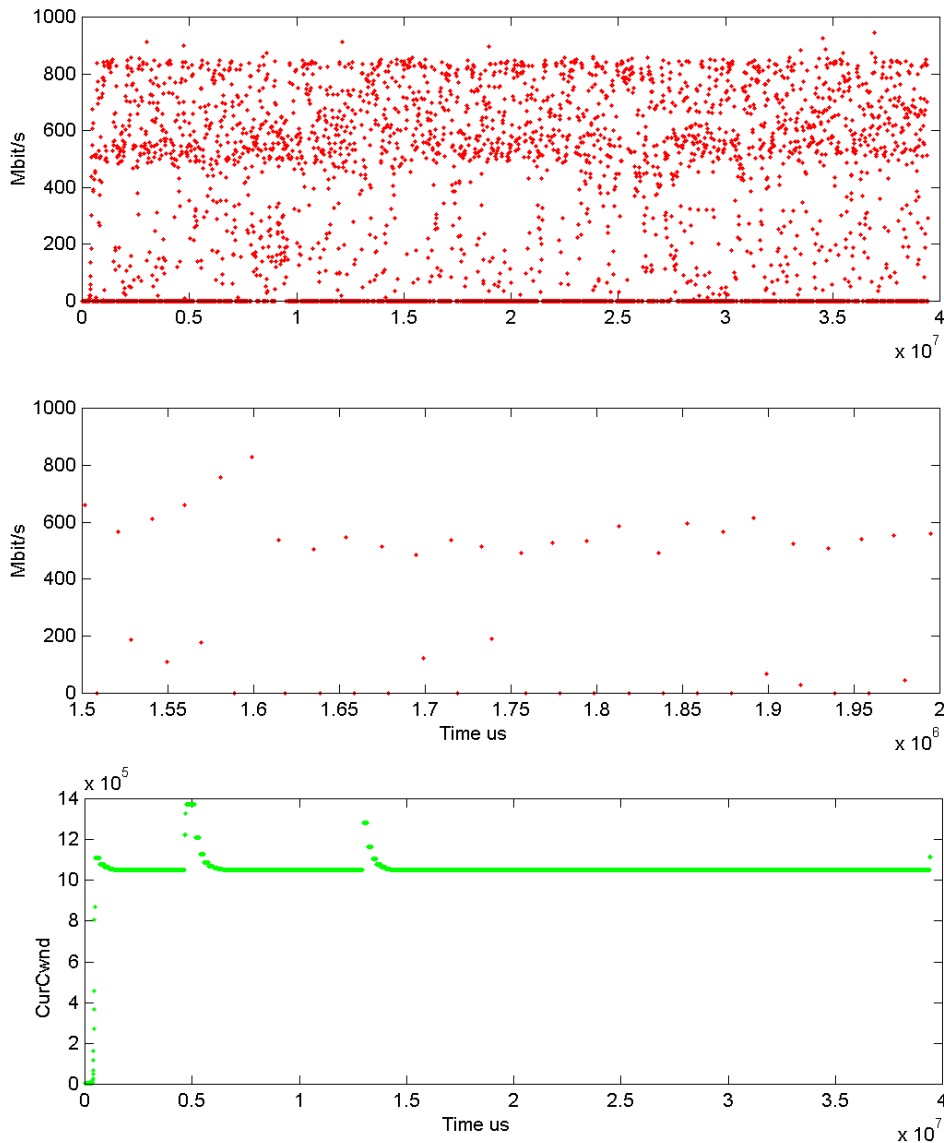


Figure 30. Time series plots of the throughput and the TCP congestion window during transfer of a 2 GByte file using GridFTP. The alternating 0 – 600 Mbit/s throughput can be seen in the middle plot.

6.7 http Data Transfers

There have been recent discussions within GridPP, DataGRID and at GGF [GGF] on the possibility of using standard Web servers e.g. apache for transferring bulk data. (Straightforward plugin extensions already allow the Grid certificates and the Grid proxy to be used for authentication.[gridsite]) In collaboration with WP6 DataGRID, throughput measurements were made using the standard apache web server and a prototype client based on the curl http library. The same hardware was used as for the GridFTP tests described above. The client first determines the size of the file from the web server, creates the file, and then transfers the data. This approach allows multiple TCP streams to be used in a simple manner. For these tests, the default TCP buffer size was set to 1M

byte - the rtt-bandwidth product. The disk-to-disk transfer rate was 72Mbytes/s and the corresponding time series throughput plots from web100 are shown in Figure 31. The top graph shows the throughput for the entire 23 seconds that it took to transfer the 2 GByte file, the average TCP throughput being ~ 725 Mbit/s. The dips in the transfer rate correspond to decreases in the advertised available space in the receive buffer or increases in the time the sender limited the transfer. The middle graph in Figure 31 shows this plot expanded for the time period 9 to 11 seconds, demonstrating the regular high throughput; and the lower plot shows the time series plot of the TCP Congestion window. After the initial slow start and the decrease of the overshoot, this shows no dramatic changes, in fact there were no lost packets during this test.

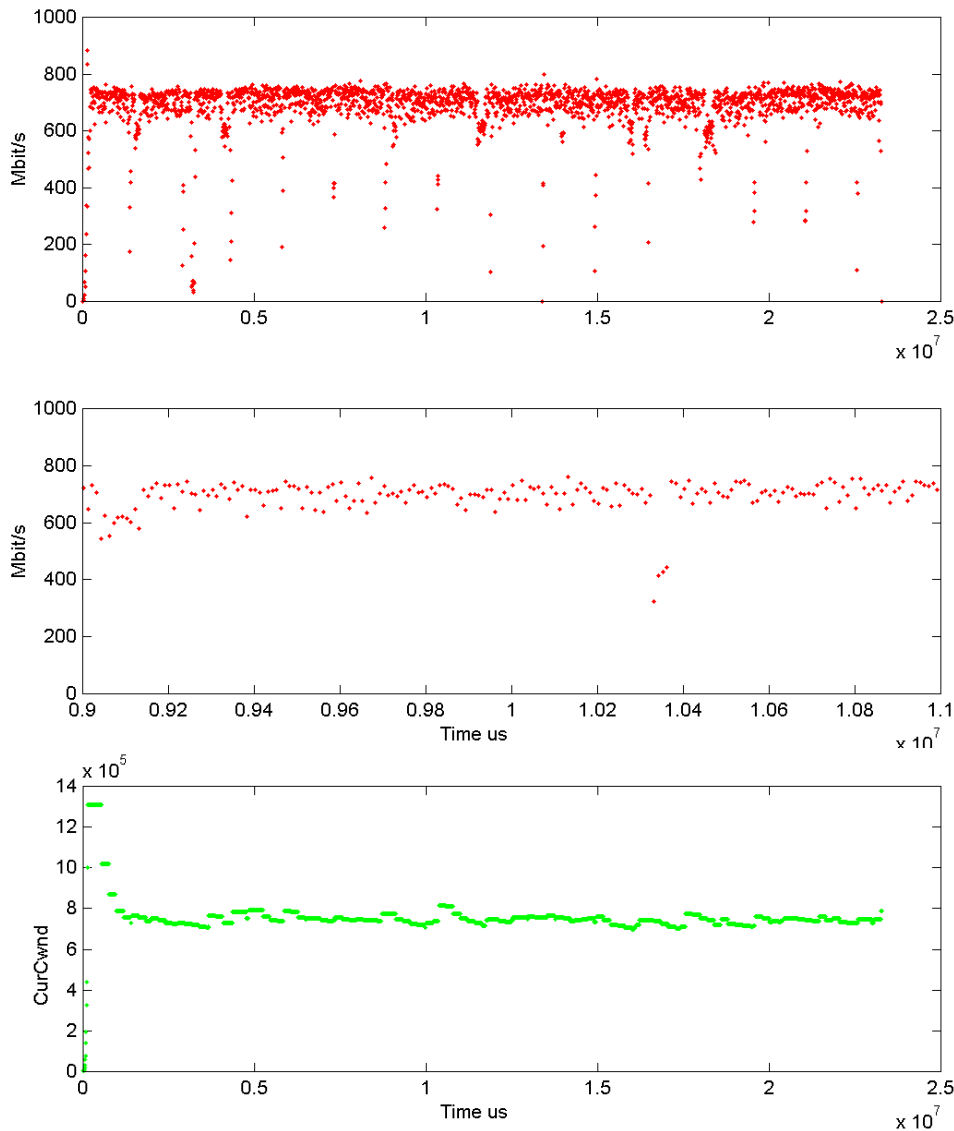


Figure 31. Time series plots of the throughput and the TCP congestion window during transfer of a 2 GByte file using http.

6.8 Transfer of Radio Astronomy VLBI Data

“Proof of concept” demonstrations at iGRID2002 23-26 September 2002 in Amsterdam, and at the European Research 2002 exhibition at the launch of the European Commission Sixth Framework Programme for research showed that the production network could be used to transfer VLBI data at over 500Mbit/s using UDP [VLBI]. These tests were made in collaboration with Dante, SURFnet and UKERNA, moving real radio astronomy data from the University of Manchester to the JIVE institute in Dwingeloo in the Netherlands. The data crossed SuperJANET4, the UK academic network, which is connected to SURFnet, the academic network in the Netherlands, by GÉANT. Experiments using the MB-NG development network are in progress to increase this to 1 Gbit/s. Recent tests from man01 to lon03 are shown in Figure 32, which shows that a user data throughput of 773Mbits/s could be achieved when the software spaced the packets at 8 μ s. During these tests just over 68.8 Million packets were sent, none were lost and none arrived out of order; this is encouraging as packet loss and corruption would not be expected on the MB-NG research network under these conditions.

Full 1500 byte Ethernet frames were used to move the data, taking into account the framing overheads, these can only be transmitted every 12.48 μ s on Gigabit Ethernet, but the application software can give data to the IP stack faster than this causing queues to build up in the sending host. The increase in throughput with packet spacing down to 8 μ s indicates an interesting interaction between reading data from disk and buffering it in the IP stack. This is being investigated further.

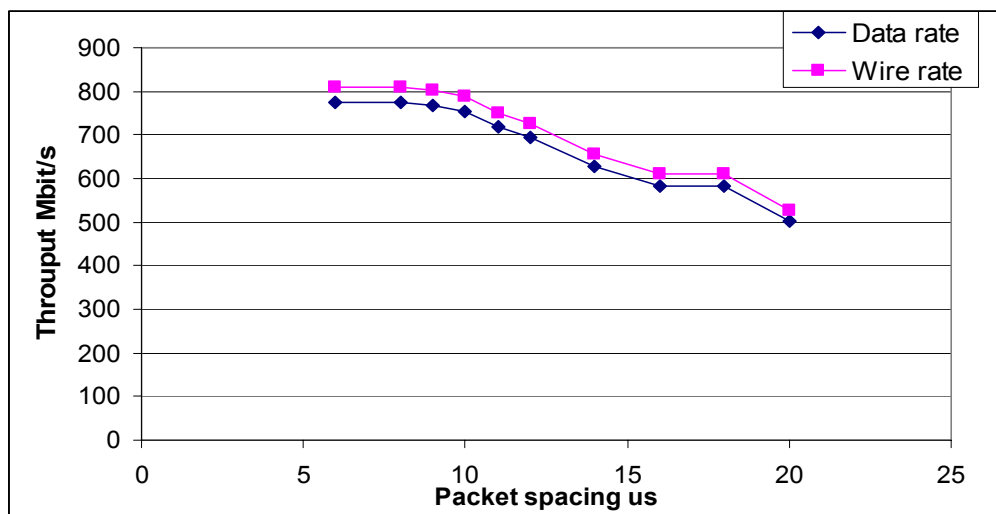


Figure 32. Transfer rates for Radio Astronomy data sent using spaced UDP packets as a function of the packet spacing. The “wire” rates include UDP, IP headers and Ethernet Framing overheads.

6.9 Conclusions

The major successes achieved the High-Bandwidth High-Throughput studies during the first year using the London - Manchester MB-NG development network were:

- Continuous memory to memory data transfers with an average user data rate of 940 Mbit/s for over 24 hours using the HighSpeed TCP stack, as shown in Figure 33.
- Sustained data transfers of 2 GByte files between RAID0 disk systems using GridFTP giving a data rate of 520 Mbit/s on the network is shown in Figure 34.
- Transfers of 2 GByte files using the http protocol from the standard apache Web server and HighSpeed TCP that achieved data rates of ~725 Mbit/s.
- Operation and comparison of different TCP stacks. More detailed study of their performance and the sharing of multiple streams will be made in the coming months.

The results show that sustained data flows at Gigabit rates are achievable, providing that “Server-quality” PCs are used and the correct selection of parameters is made.

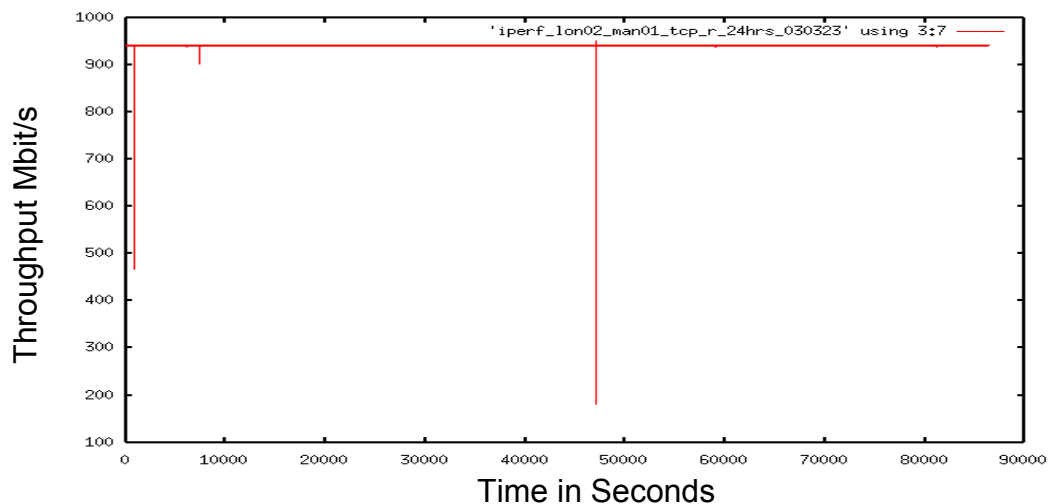


Figure 33. 24 Hour time series throughput for memory-memory transfers with 1 stream using the HighSpeed TCP stack.

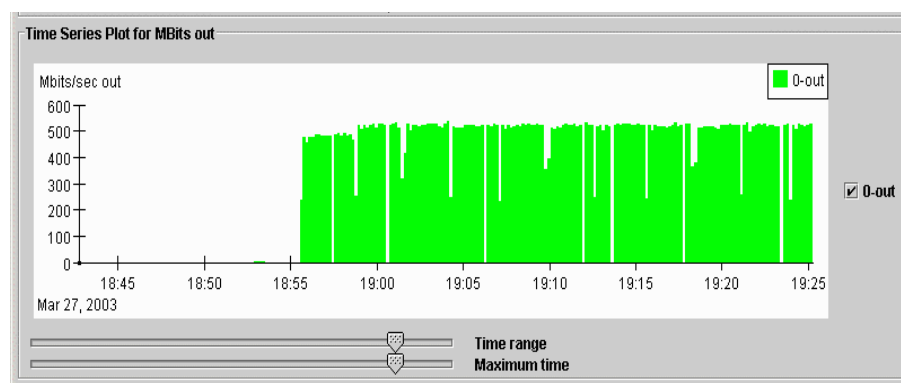


Figure 34. 24 Hour time series throughput for memory-memory transfers with 1 stream using the HighSpeed TCP stack.

References

- [UDPmon] Source and Documentation available at www.hep.man.ac.uk/~rich/net .
- [3ware] 7500-8 Benchmarks <http://www.3ware.com/products/benchmarks.asp>
- [HSTCP] The HighSpeed TCP stack and some results are available from <http://www.hep.man.ac.uk/u/garethf/hstcp/>
- [ScalableTCP] A discussion of the Scalable TCP stack and the patch are available from <http://www-lce.eng.cam.ac.uk/%7Ectk21/scalable/>
- [GGF]The Data Transport Research Group <http://www.epcc.ed.ac.uk/~dtrg/meetings.php#GGF7>
- [gridsite] Discussion of the plugin extensions may be found at: www.gridpp.ac.uk/authz/gridsite
- [VLBI] See www.hep.man.ac.uk/~rich/VLBI_web for further information on the iGrid2002 and ER2002 demonstrations and the paper "High Data Rate Transmission in High Resolution Radio Astronomy - vlbiGRID" http://www.hep.man.ac.uk/u/rich/VLBI_web/igrid2002_astro/iGrid2002_v8.pdf