DataTAG

HIGH PERFORMANCE NETWORKING

END TO END INTER-DOMAIN QOS

Document identifier:	**DATATAG-D2.3-Final-last**
EDMS id:	**431720**
Date:	**17/02/2004**
Work package:	**WP2**
Partner(s):	**CERN, INFN,INRIA,PPARC,UvA**
Lead Partner:	**PPARC**
Document status :	**DRAFT**
Deliverable identifier:	**D2.3**

Disclaimer : The results presented in this document must be considered exclusively in the scope of the programme of work of the DataTAG project - QoS implementations by router vendors may not be evaluated beyond the exact hardware and software used to run the tests

Abstract: This document is the final deliverable of DataTAG WP2 Task 2.2 and describes experimental work to investigate the properties of end-to-end inter-domain QoS as delivered by routers running Differentiated Services. Tests have been performed using both 1 GigE and OC-48 line cards from three router vendors and both absolute and comparative studies have been made. From a more experimental perspective Equivalent Differentiated Services are also investigated running in software routers and promising results are reported.

Delivery Slip

	Name	Partner	Date
From	Robin Tasker	PPARC	16-12-2003
Reviewed by	Olivier Martin, Elise Guyot	CERN	18-12-2003
Approved by	Olivier Martin	CERN	19-12-2003

Document Log

Issue	Date	Comment	Author
1	8-12-2003	Initial draft	Robin Tasker
2	15-12-2003	First Draft	Robin Tasker
3	16-12-2003	Second Draft	Robin Tasker
4	17-2-2004	Final draft	Robin Tasker

Document Change Record

Issue	Item	Reason for Change
1	Initial draft	Tasker
2	1. DS	Di Donato
	2. EDS	Primet
3	3. Editorial improvements 1. Minor corrections 2. Editorial work	Tasker Multiple sources Tasker
4	1. Disclaimer added, some inaccuracies about the PRocket router evaluation fixed.	Di Donato &Tasker & Procket

Files

Software Products	User files / URL
Word	D2.3-Final-last.doc

CONTENT

1 INTRODUCTION 4

1.1 OBJECTIVES OF THIS DOCUMENT 4
1.2 APPLICATION AREA 4
1.3 APPLICABLE DOCUMENTS AND REFERENCE DOCUMENTS 4
1.4 TERMINOLOGY 5
1.5 ACKNOWLEDGEMENTS 6

2 EXECUTIVE SUMMARY 7

3 DIFFERENTIATED SERVICES 8

3.1 INTRODUCTION 8
3.2 OBJECTIVES 8
3.3 EXPERIMENTAL DESIGN - BACKGROUND TRAFFIC 8
3.4 EXPERIMENTAL DESIGN - QOS TECHNIQUES INVESTIGATED 9
3.5 TEST METHODOLOGY 9
3.6 ERROR ANALYSIS 10
3.7 PER ROUTER-MANUFACTURER TEST RESULTS 11
3.7.1 Cisco 11
3.7.2 Juniper 18
3.7.3 Procket 25
3.8 COMPARATIVE ANALYSIS 29
3.8.1 M-LREWS (Max LBE Relative Error With Sign) 29
3.8.2 A-ALRE (Average Absolute LBE Relative Error) 31
3.9 CONCLUSIONS 33
3.10 FUTURE WORK - SYNTHESIS 35

4 EQUIVALENT DIFFERENTIATED SERVICES 36

4.1 INTRODUCTION 36
4.2 PER HOP BEHAVIOUR - IMPLEMENTATION AND IMPROVEMENT 36
4.3 EDS TRANSPORT LAYER DESIGN 37
4.3.1 End-to-end delay constrained transport protocol over EDS:RT-TP 38
4.3.2 Interactive reliable transport protocol over EDS: SM-TP 38
4.3.3 Bulk reliable transport protocol over EDS: LM-TP 39
4.4 LM-TP OVER EDS IMPLEMENTATION AND EVALUATION 39
4.4.1 Tests Methodology 39
4.5 LM-TP EVALUATION 39
4.6 CONCLUSIONS 41

1 INTRODUCTION

1.1 OBJECTIVES OF THIS DOCUMENT

This document is the final deliverable of Task 2.2 of DataTAG WP2, End-to-End Inter-Domain QoS. It develops the work introduced in the first DataTAG WP2 deliverable, D2.1 [Deliverable D2.1] in both the areas of differentiated services [DS] and equivalent differentiated services [Montenegro].

The work of differentiated services examines the capabilities of both OC-48 and GigE line card interfaces from three router vendors in some detail, and describes the start of work to investigate the use of differentiated services running over the variant transport protocols investigated by DataTAG WP2, Task 2.1.

Equivalent differentiated services are more experimental and the work described here is in way a proof of concept and a pointer towards further work. Undoubtedly the opportunity to work within the DataTAG environment has been of considerable benefit when progressing these ideas and concepts.

1.2 APPLICATION AREA

QoS is a much discussed topic, with many examples of performance improvements in development or testbed networks, but little wide-scale deployment in service networks. More recently, the focus of attention has been on the Less than Best Effort (LBE) services and the benefits that such an approach provides. This fits well with expected Grid applications where persistent data transfers will be required. It is less clear where the need for more timely data arrival will be required, although work to provide enhanced IP services (e.g., Premium IP) is underway within the provider networks.

DataTAG WP2 is also reviewing the alternatives to the conventional approaches to differentiated services, specifically Equivalent Differentiated Services.

1.3 APPLICABLE DOCUMENTS AND REFERENCE DOCUMENTS

Reference documents

Dovrolis	Dovrolis and Ramanathan (1999) A case for relative differentiated services and the proportional differentiated model *IEEE Network*, 13(5): 26-34
DS	Differentiated Service: RFCs 2475, 2597, 2598 and 3246
Floyd	1. High Speed TCP for Large Congestion Windows, Sally Floyd., Internet draft draft-floyd-tcp-highspeed-01.txt, work in progress, 2002.
	2. Limited Slow-Start for TCP with Large Congestion Windows Sally Floyd, Internet draft draft-floyd-tcp-slowstart-01.txt, work in progress, August 2002.
Goutelle	Goutelle, Gadioz and Primet (2002) Resultats preliminaires sur le comportement de TCP au dessus d'une couche a services differencies equivalents *Technical Report RR-4634*, INRIA
Hurley	Hurley, LeBoudec, Thiran and Kara (2001) ABE: Providing a low-delay service within best effort. *IEEE Network*, 15(5):60-69
Kelly	1. On engineering a stable and scalable TCP variant., Cambridge University Engineering Department Technical Report CUED/F-INFENG/TR.435, June 2002.
	2. Scalable TCP: Improving Performance in Highspeed Wide Area Networks, Submitted for publication, December 2002.
Low	1. A new TCP/AQM for Stability and Performance in Fast Networks , Fernando

Paganini, Zhikui Wang, Steven H. Low, John Doyle, Proc. of 39th Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, October 2002

2. http://netlab.caltech.edu/FAST/ and

3. http://www-iepm.slac.stanford.edu/monitoring/bulk/fast/

Montenegro — Montenegro, Gadioz, Primet and Tourancheau (2002) Equivalent differentiated services for AOD-Vng *ACM SIGMOBILE Mobile Computing and Communications Review*, 6(3):110-111

QoS1 — White Paper -The Need for QoS. The Internet Protocol's "best-effort" service has worked well so far, so why do we need to change it?

http://www.qosforum.com/

QoS2 — White Paper - QoS protocols & architectures Quality of Service protocols use a variety of complementary mechanisms to enable deterministic end-to-end data delivery

http://www.qosforum.com/

QoS3 — White Paper - Introduction to QoS Policies
Quality of Service protocols provide the mechanics to differentiate traffic, and Policy defines how they're used.
http://www.qosforum.com/

RED — Floyd and Van Jacobson (1993) Random early detection gateways for congestion avoidance. *IEEE/ACM Transactions on Networking*, 1(4):297-413

Teitelbaum — Future Priorities for Internet2 QoS, October 2, 2001

1.4 TERMINOLOGY

Definitions

API	Application Programming Interface
BE	Best-effort
CBR	Constant Bit Rate
CIR	Committed Information Rate
DA	Destination Address
DRR	Deficit Round Robin
DS	Differentiated Services
EDG	European Data Grid
EF	Expedited Forwarding
FTP	File Transfer Protocol
Globus	See http://www.globus.org
GridFTP	Grid File Transfer Protocol

HEP	High Energy Physics
HTML	Hypertext Markup Language
LBE	Less-than-Best-Effort
IETF	Internet Engineering Task Force
Linux	An open source variant of Unix
MRTG	Multi Router Traffic Grapher
MTU	Message Transfer Unit
NTP	Network Time Protocol
PHB	Per Hop Behaviour
PIR	Peak Information Rate
PLR	Proportional Loss Rate
QoS	Quality of Service
QBSS	QBONE Scavenger Service
RED	Random Early Discard
RFC	Request For Comment, a formal document of the IETF
RTO	Retransmit TimeOut
SA	Source Address
SLA	Service Level Agreement
SNMP	Simple Network Management Protocol
TCP	Transmission Control Protocol
TOS	Type of Service, a field within the IP header of IP datagrams.
UDP	User Datagram Protocol
Unix	A generic computer operating system
URL	Universal Resource Locator
WFQ	Weighted Fair Queue
WRR	Weighted Round Robin
WTP	Wait Tail Priority

1.5 ACKNOWLEDGEMENTS

The work of WP2 Task 2.3 greatly benefited from the generous support of the US Department of Energy (DoE) through the California Institute of Technology (CALTECH) who contributed manpower and hardware resources to the DataTAG project.

2 EXECUTIVE SUMMARY

This document is the final delivery from DataTAG WP2 Task 2.2, End-to-End Inter-Domain QoS and reports on the work carried out up to December 2003. The DataTAG project has been extended until March 2004 and QoS work will continue throughout this extension period with the formal presentation of the additional work at the final EU Review. However an outline of this work is provided here.

The work described here can be divided into two distinct areas which follow on from the preliminary experimental work presented in DataTAG WP2 Deliverable D2.1 in December 2002. Work has therefore been carried out on traditional *differentiated services* and on *equivalent differentiated services*.

Traditional *differentiated services* are generally well supported by router vendors and it was the capabilities of these implementations that have been the subject of investigations described here. An important component of this work is the synthesis of differentiated services operating over the new TCP stacks that have been investigated as a part of the work of DataTAG WP2 Task 2.1. It is this second component that will be reported in full in March 2004; here the initial finding are described.

At the time of the first Task Group deliverable, the availability of the DataTAG provision coupled with problems with the QoS implementations themselves within the routers under test had limited the extent of the results presented. Happily those problems have been overcome and, with the addition of routers from Juniper and Procket, a full range of QoS characterisation tests have been performed on both the OC-48 and the GigE line card interfaces. This characterisation of the router line cards provides essential information for the correct engineering of IP-level bandwidth allocations within a network.

In all cases the tests described here were carried out across the DataTAG provision between CERN and Chicago and have made use of the DataTAG test environment for the generation of both test and background traffic. The focus of the work has been to understand the operation and inter-working between the differentiated service classes known as Best Effort (BE) and Less-than-Best Effort (LBE) modelled in the routers using bandwidth scheduling algorithms known as Weighted Fair Queuing (WFQ), Weighted Round Robin (WRR) and Deficit Weighted Round Robin (DWRR).

From the tests it is apparent that the QoS implementations in the OC-48 line cards are more precisely formulated than that found for the GigE line cards. However differences exist between the vendors' implementations for both GigE and OC-48 line cards and these differences are discussed.

The study has highlighted how good link utilisation is necessary but not the only determinant for precise bandwidth allocation between the differentiated services classes of BE and LBE. The study has shown that router response to the proportion of traffic configured to the different classes and the level of per-port congestion is also of significance.

Equivalent differentiated services (EDS) are experimental and as such are not as yet supported by the router vendors. These services represent a radical departure from the traditional approach insofar that they provide a range of "different but equivalent" services that are a trade-off between delay and of loss rates for the end-to-end flow. In this respect EDS is broadly analogous to TCP operating over IP where the transport protocol has to perform some flow adaptation. The aim and purpose of EDS is to improve the overall global end-to-end network performance and not to improve the performance of each separate co-operative flow.

This work has included the implementation of the EDS principle within software routers for back-to-back testing in the first instance and subsequently across the wide area and the higher data rates available in the DataTAG provision. Such software routers are straightforward to deploy at the network edge, and as the EDS does not rely upon the bounded domain nor on the pricing concepts associated with traditional differentiated services, they are able to mitigate against performance bottlenecks found so often at the point of ingress into a network.

3 DIFFERENTIATED SERVICES

3.1 INTRODUCTION

For the successful deployment of IP QoS based upon the use of Differentiated Services (DS) [DS] in both access and core networks, an understanding of the performance of both the 1GigE (1Gbps) and the POS/OC48 router interfaces (line cards) is required. Such an evaluation of performance is based upon how precisely a minimum bandwidth guarantee can be allocated to an aggregate of data traffic under interface congestion. This information constitutes the foundations for the deployment of more complex IP QoS solutions.

The QoS performance of these interfaces has been investigated for Cisco, Juniper and Procket routers such that a comparison may be made with respect to their performance. In addition such investigation will inform the QoS engineering process of a multi-vendor network.

For the deployment of QoS and defining sensible Service Level Specification (SLS) and Service Level Agreement (SLA), it is important that the network behaviour is quantified and understood. The QoS model used here is based on the DS model for IP networks. Traffic entering the network device is marked using a single Differentiated Services Code Point (DSCP) and for each one of these code points there is assigned a different behaviour aggregate or class.

3.2 OBJECTIVES

The principle objective of the tests was to understand the line card behaviour in order to inform the correct engineering of IP-level bandwidth allocation.

The approach taken was to look at the utilisation of the link, e.g. the proportion of the available link capacity that is actually used, together with both the absolute and the relative bandwidth allocation errors (See 3.6 for details). This double metric was necessary so that a complete picture of the performances of the routers under test could be developed.

In particular, a bad link utilisation provides a quick and precise method to detect anomalies but of course integrated with a proper error analysis. The latter is useful to detect the presence of errors when correctly allocating bandwidth even where the link utilisation is good. It is also useful to determine where the errors are localised when the link utilisation shows poor performances.

3.3 EXPERIMENTAL DESIGN - BACKGROUND TRAFFIC

In the following discussion the router under test will be referred by name, i.e. "Cisco", "Juniper", etc, while the generic router needed to connect each data source with its own data sink will be referred to as “GR”.

Three PCs (Supermicro 6022P-6 Dual Intel® Xeon) were attached to the two routers. Each PC had an Intel® PRO/1000 XT Server Gigethernet adapter (e1000 v4.4.12-k1) with the PCs running Linux kernel version 2.4.20. The two routers were connected back-to-back, for example. "Juniper" to “GR”, where the "GR" is not changed between tests. The routers were connected using either POS OC-48 (2.5Gbps) or GigE (1Gbps Ethernet) line cards.

To baseline the performance of the PCs, two were connected back-to-back and the throughput versus packet size was measured. The results of these tests are shown in Figure 1.

To generate traffic from the PCs iperf (version 1.6.5 (13 Jan 20030) pthreads) was used to produce a constant bit rate (CBR) pattern with UDP used as the transport protocol.

From Figure 1 it can be seen that to achieve line rate from the PCs, a packet size quite close to

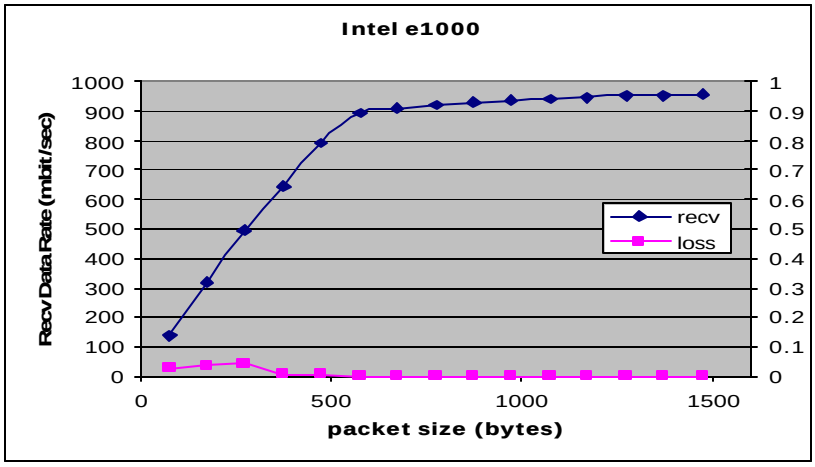

Figure 1 : Back-to-back PC performance measured as received data rate as a function of packet size

the Ethernet MTU was required. A packet size of 1470 bytes was chosen for the tests where the maximum achieved throughput at this packet size for the PCs plugged back-to-back was 955Mbps.

3.4 EXPERIMENTAL DESIGN - QOS TECHNIQUES INVESTIGATED

The main aim of the tests was to allocate different proportions of the bottleneck bandwidth to two different DS classes, namely Best Effort (BE) and Less Than Best Effort (LBE). In this respect the use of LBE is equivalent to the Scavenger service defined by Internet2.

BE class was defined with DSCP=0 and LBE class with DSCP=8 (001000) which is consistent with the recommendation from Internet2. It should be noted that for the tests described here the packets were marked with the DSCP code at the PCs before being transmitted.

The bandwidth scheduling algorithm available on the three routers differed. Cisco routers use the algorithm known as Weighted Fair Queuing (WFQ), Juniper routers use Weighted Round Robin (WRR) while Procket routers use Deficit Weighted Round Robin (DWRR). The traffic type used to investigate the performance of the bandwidth scheduler algorithms (WFQ, WRR and DWRR) was in all cases UDP.

3.5 TEST METHODOLOGY

With the overall number of injected flows kept constant at either three or two, three major tests were conducted:

- **Test1** where 2 BE flows and 1 LBE flow were injected;
- **Test2** where 1 BE flow and 2 LBE flows were injected; and
- **Test3** where 1 BE flow and 1 LBE flow were injected.

The offered load of each one of the flows ranged from 100Mbps to 1Gbps, with the highest granularity possible compatible with the combined performances of both the iperf tool and Linux PCs used to send traffic.

The bandwidth allocation between BE and LBE classes chosen for the tests was as follows:

BE-LBE = (99-1, 98-2, 97-3, 96-4, 97-3, 95-5, 94-6, 93-7, 92-8, 91-9, 90-10, 85-15, 80-20, 75-25, 70-30, 65-35, 60-40, 55-45, 50-50)

In all results presented here, the sequence of BE-LBE allocations shown above is described as the "Bandwidth Allocation Couples Axis" with the axis direction from 99-1 to 50-50.

Measurements were taken of atomic metrics such as the per-class received throughput and packet loss. However a composed metric for the evaluation of the precision by which the allocation of bandwidth occurred amongst the classes involved was defined and used along with the link utilisation metrics.

3.6 ERROR ANALYSIS

Bad link utilisation is only one reason for bad bandwidth allocation and therefore errors in allocating bandwidth are also shown. In particular, two types of errors, relative and absolute, are seen as a measure of the precision in the allocation of bandwidth for both the BE and LBE class.

The absolute (Mbps) and relative (%) errors for a generic class X are:

$$AbsoluteError = WhatclassX\ gets - WhatclassX\ shouldget \quad \text{(Mbps)}$$

$$Relative Error = [AbsoluteError] * 100 / WhatClassX\ shouldget \quad (\%)$$

Note that both the Absolute and Relative error can assume an algebraic value. A negative or a positive error means that the generic Class X has been allocated more or less than it should have been allocated respectively.

The two formulae deriving the absolute and relative error rely on the computation of the parameter "What Class X Should Get". The following schema (Figure 2) shows the algorithm used to work out the expected received throughput when the number of classes in the system is two. Such schema scales to any number of classes employed.

This algorithm is based on the (Class Based) Weighted Fair Queuing (CBWFQ) algorithm which governs the logic of all the three different manufacturer's line cards under test.

It is worth noting that most of the manufacturers fix the accuracy in allocating bandwidth in a class to be around 95% which in turn means that 2.5% is the max error acceptable. We refer to the region inside which this bound is validated as the "operating region".

In order to bound an operating region from the bandwidth allocation couples axis, a new metric consisting of the maximum value that the LBE and BE relative errors with sign, assume over the offered load axis, is introduced. The errors are taken into account with their sign as the polarity of the error is functional to an understanding its nature. These metrics are referred to as the "Maximum LBE Relative error With Sign" (M-LREWS) and the "Maximum BE Relative error With Sign" (M-BREWS) respectively.

The MAX LBE/BE Relative Errors with sign (M-LREWS/M-BREWS) allow the evaluation of the bandwidth scheduler based solely on its worst performance over the (offered load/port congestion level) axis.

This is of extreme importance since the bounded value for the precision in the allocation of bandwidth to which the manufacturers refer can be correctly associated to the worst case scenario out of the

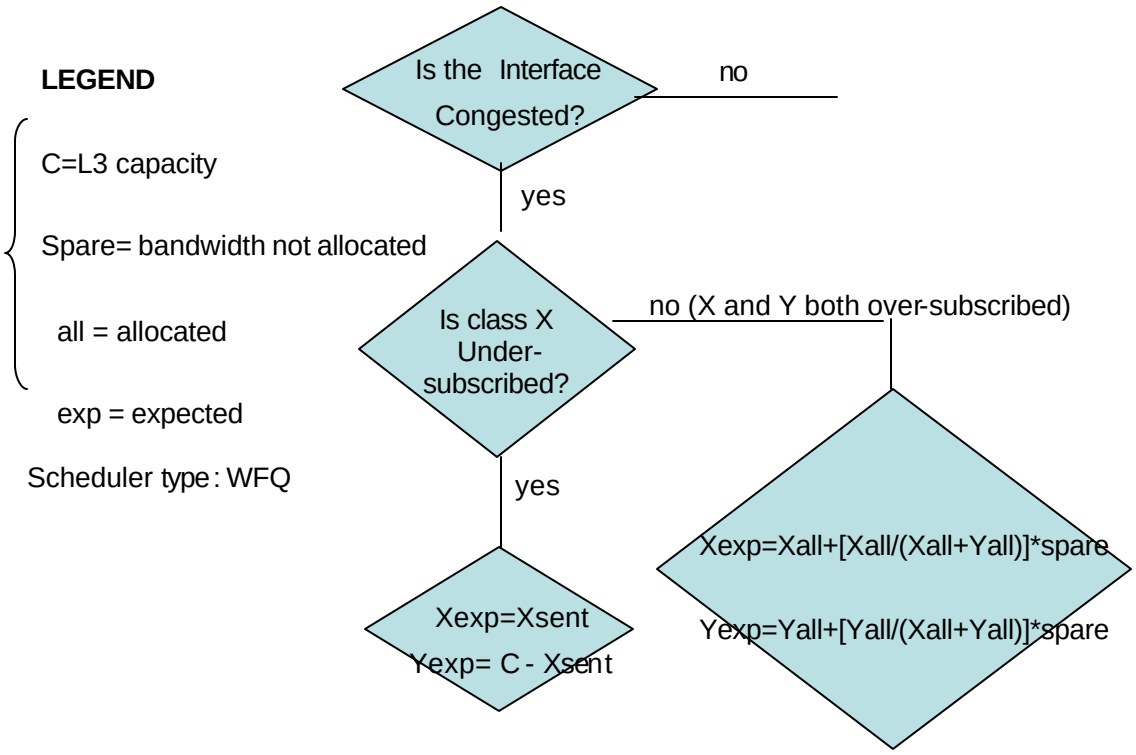

Figure 2 : Schema showing the algorithm used to work out the expected received throughput when the number of classes in the system is two

whole offered load axis. It is therefore correct to say that an accuracy of 95% in the allocation of the bandwidth is equivalent to having the M-LREWS/M-BREWS <= +-2.5%.

We refer to this operating region as the "max operating region", thus highlighting that the method used to bound it was that of computing the max algebra for the relative errors.

In order to evaluate the performances of a line card averaged over the whole offered load axis, or better, averaged over different card congestion levels, another metric consisting of the average of the absolute values of LBE/BE relative errors in introduced. This is referred to as A-ALRE and A-ABRE for LBE and BE respectively. The absolute values are used here to avoid the situation where the average of the algebraic values could lead to misleading ~0% errors.

The main difference between the MAX-based and the Averaged-based metrics is that the latter takes account of the errors over all the "offered_load" / "card_congestion_level" axis and not just of the maximum. This allows the spread of the error over the card congestion level axis to be quantified.

3.7 PER ROUTER-MANUFACTURER TEST RESULTS

3.7.1 Cisco

3.7.1.1 Testbed

Figure 3 outlines the layout of the test-bed for the OC-48 line card tests.

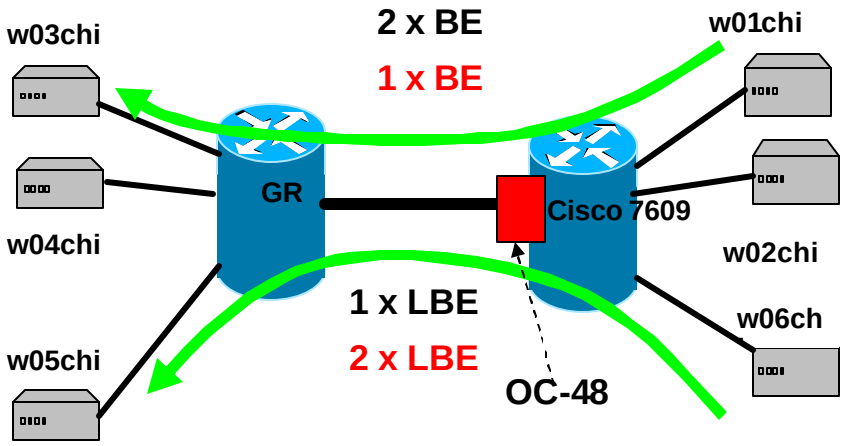

Figure 3 : The layout of the test-bed for the Cisco OC-48 line card tests

The line card under test was a POS OC-48 v2 (referred to by Cisco as OSM-1OC48-POS-SS+) with the encapsulation used is PPP.

Cisco designed an "engineering code" specific for the scheduler of this card and included it on the major release 12.1(19)E which was available from May 2003.

Figure 4 shows the test-bed layout for the Cisco GigE-WAN line card tests.

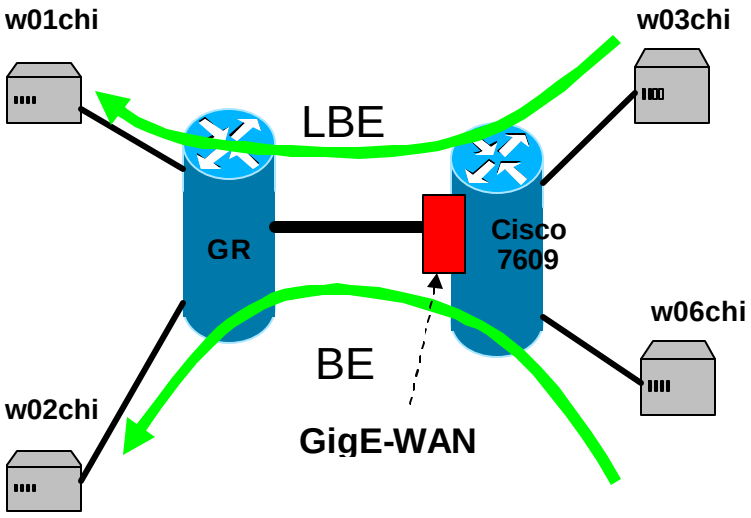

Figure 4: The layout the test-bed layout for the Cisco GigE-WAN line card tests.

This line card under test was a GigE-WAN v2 (referred to by Cisco as OSM-2+4GigE-WAN+). The tests were again carried out using the 12.1(19)E IOS version, i.e. the same as used for the OC-48 line card test.

3.7.1.2 OC-48 and 1GigE Configuration

The recommended configurations used for the OC-48 and for the GigE-WAN was as follows,

```
!
class-map match-any BE
  match ip dscp 0
class-map match-any LBE
  match ip dscp 8
!
policy-map UCL
  class BE
    bandwidth percent X
  class LBE
    bandwidth percent Y
!
mls qos
!
interface input
mls qos trust dscp
!
interface output
service-policy output ucl
mls qos trust dscp
!
```

[Note: "mls qos" in the global configuration mode was needed to enable QoS on the supervisor engine while "mls qos trust dscp" issued in the input and output interfaces was there to avoid cards resetting the dscp code of packets entering or leaving the interfaces. This configuration line was of particular importance where Catalyst cards were used in the input (but not in the output as they do not support L3 CBWFQ) as they naturally tended to reset to 0 the dscp code. This happens because the legacy L2 COS-based QoS was the default QoS for the Catalyst ports and the 7600 router has been built on top of the native Catalyst switch.

Cisco "Modular Quality of Service Command" (MQC) for both OC-48 and GigE-WAN can be used. This is of particular importance when configuring QoS on Cisco cards/routers implementing bandwidth allocation through different algorithms such as CB-WFQ and MDRR (the latter being implemented in Cisco 12000 platforms) as MQC leaves such bandwidth scheduler implementations aside.

The max of the sum of X and Y present in the configuration cannot be set to 100% as 1% is always made available to host routing updates and network control traffic in general. Therefore (X+Y) <= 99%. The difference |100-(X+Y)| is what is called "spare" in the error analysis paragraph.

As an architectural note, Parallel Express Forwarding (PXF) was present on each OSM (Optical Service Module) or card and is capable of CBWFQ, thus permitting the QoS processing to be performed directly on the card.]

3.7.1.3 OC-48 Results

Figure 5 shows the link utilisation as a function of the per-bandwidth allocation couple. The iperf UDP-payload-level capacity C of the link, which was obtained by congesting the interface and not configuring QoS was 2318 Mbps; The card was therefore congested up to (957*3)/2318=123.8% of its capacity.

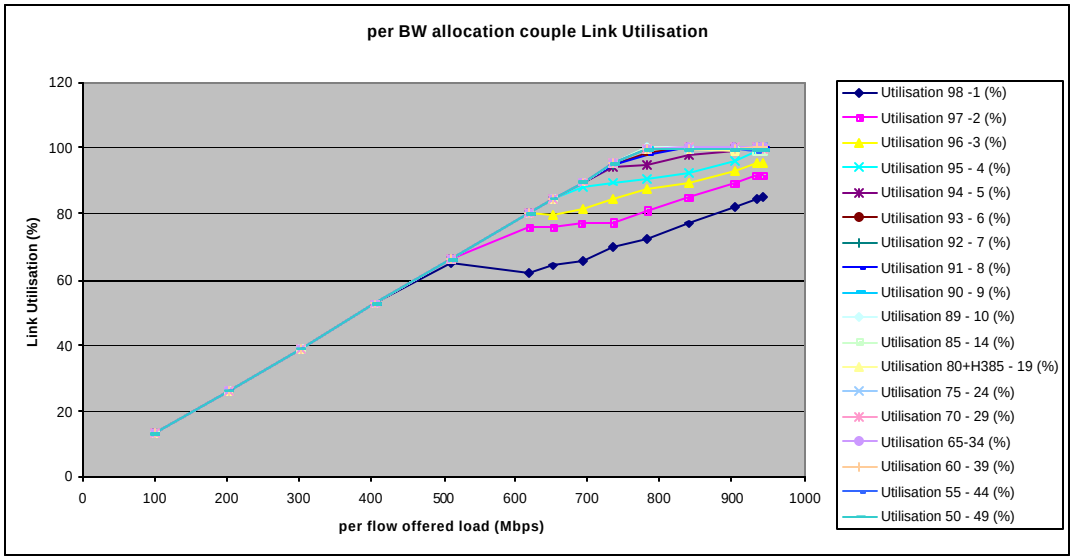

Figure 5 shows the link utilisation as a function of the per-bandwidth allocation couple

Figure 5 further shows how the utilisation of the link decreases as the bandwidth allocation couple axis

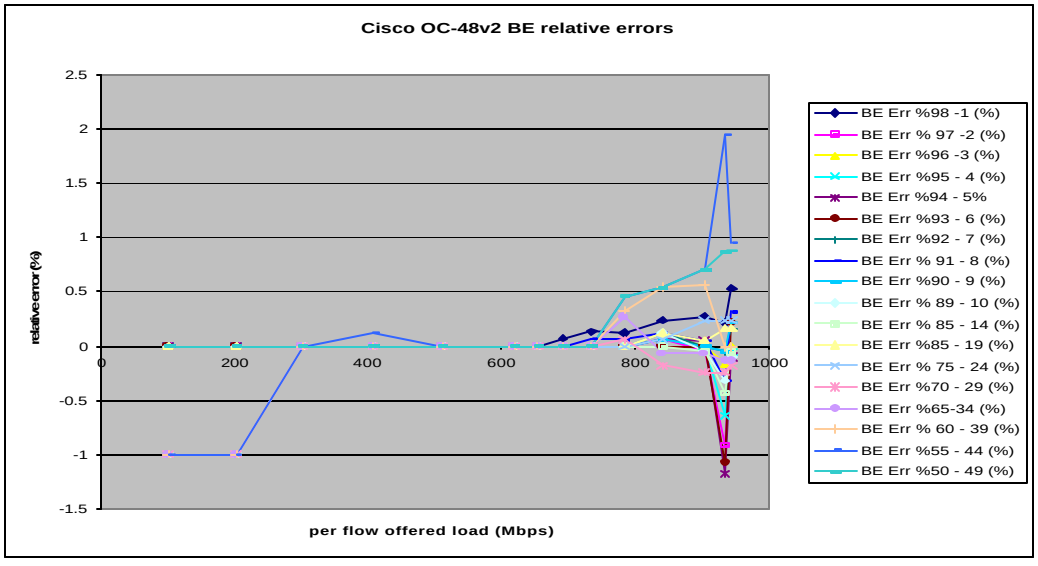

Figure 6 : The per-bandwidth allocation couple relative errors against the port congestion level for BE

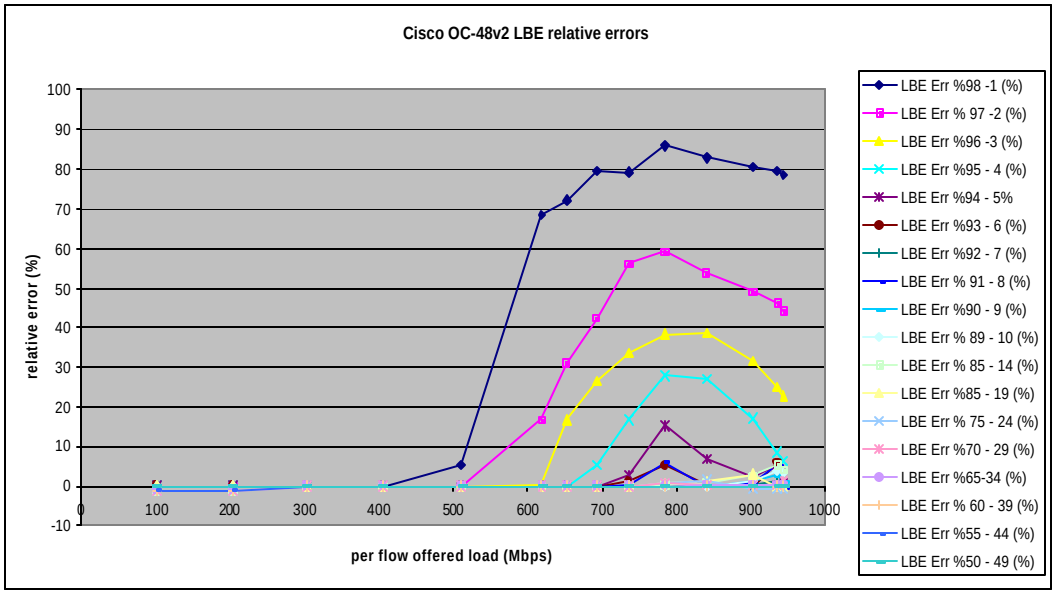

Figure 7 : The per-bandwidth allocation couple relative errors against the port congestion level for LBE

decreases. As poor link utilization is sufficient for having bad bandwidth allocation precision, the per-bandwidth allocation couple relative errors against the port congestion level for both BE and LBE are shown (in Figure 6 and Figure 7 respectively) with the purpose of quantifying the errors and localizing which card congestion level region the errors span.

The error BE presents is < 2% and therefore negligible, however the error is concentrated on LBE and presents positive polarity which suggests, along with the negligible BE error and with the poor link utilization, that the scheduler's was deficient in allocating the BE leftover bandwidth to LBE under a certain range of port congestion levels.

For a certain level of port congestion, this error decreases monotonically with the increase of the bandwidth allocation couple axis, and therefore suggests a well defined operating region.

In order to determine with precision the operating region, the MAX LBE relative error with sign (M-LREWS) is presented in Figure 8

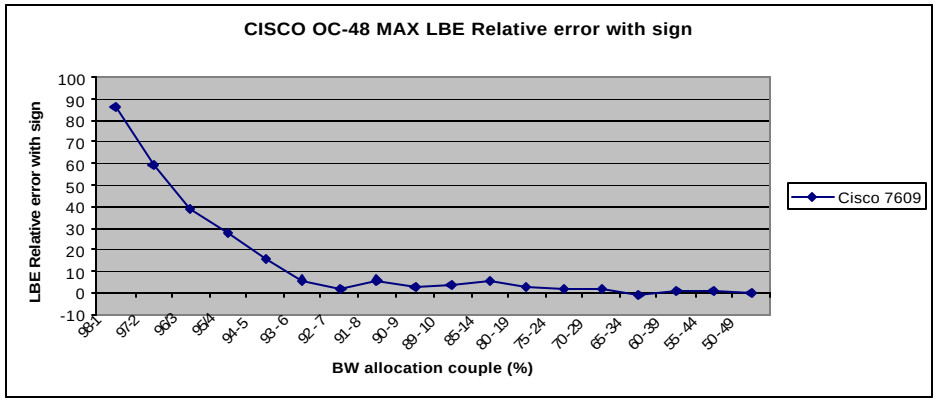

Figure 8 : The MAX LBE relative error with sign (M-LREWS) for the OC-48 line card

The error oscillates somewhat along the value of +2.5%, thus making the definition of the "max" operating region difficult. A conservative "max" operating region for this card would therefore appear to be from the value of 50-49 to that of 75-24 for the bandwidth allocation couples. The same "max" operating region would range from 50-49 to 93-6 if the precision was 88% instead of 95%.

3.7.1.4 1GigE-WANv2 results

Figure 9 shows the per-bandwidth allocation couple link utilisation against increasing card congestion levels for the 1GigE WAN line card.

The iperf UDP-payload-level capacity of the link, which was obtained by congesting the interface and not configuring QoS, was 957 Mbps; The card was therefore congested up to (957*2)/957=200% of its capacity.

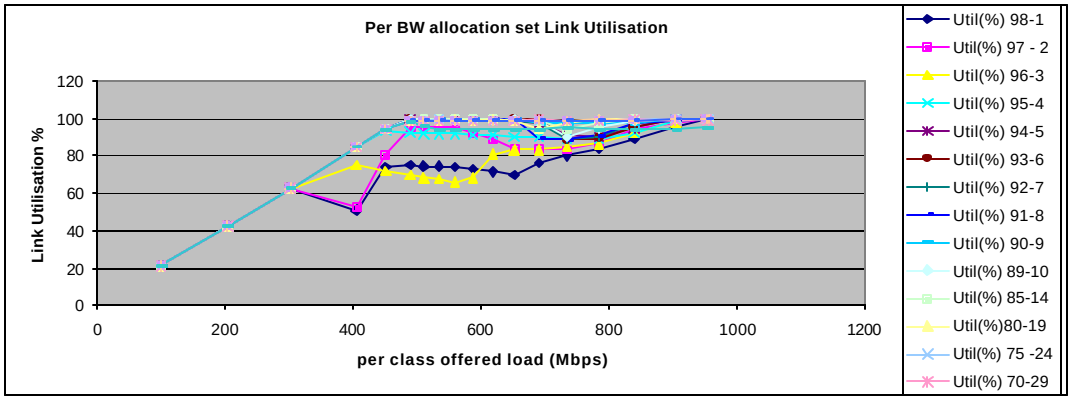

Figure 9 : The per-bandwidth allocation couple link utilisation against increasing card congestion levels for the 1GigE WAN line card.

The link utilization for this line card was poor, sufficiently so to demonstrate poor bandwidth allocation precision. Both BE and LBE relative errors are therefore presented in Figures 10 and

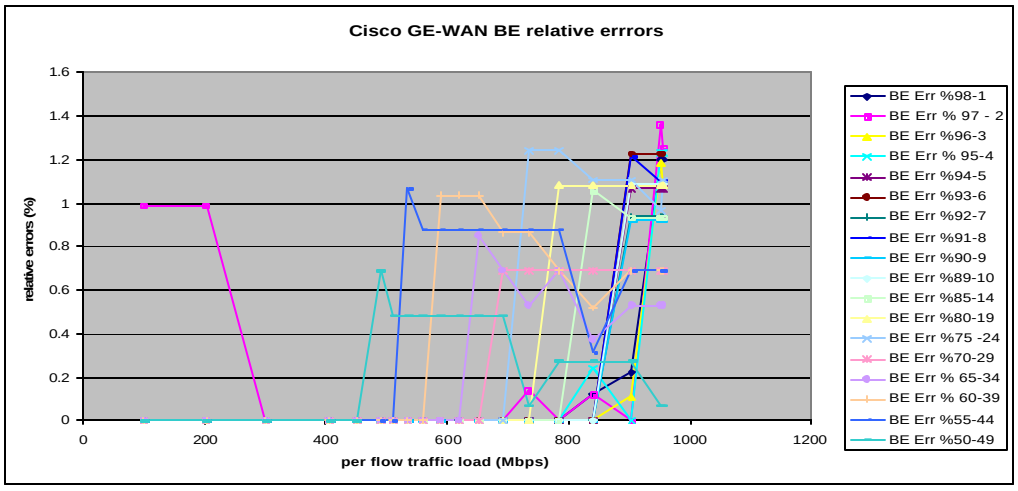

Figure 10 : The per-bandwidth allocation couple relative errors against the port congestion level for BE

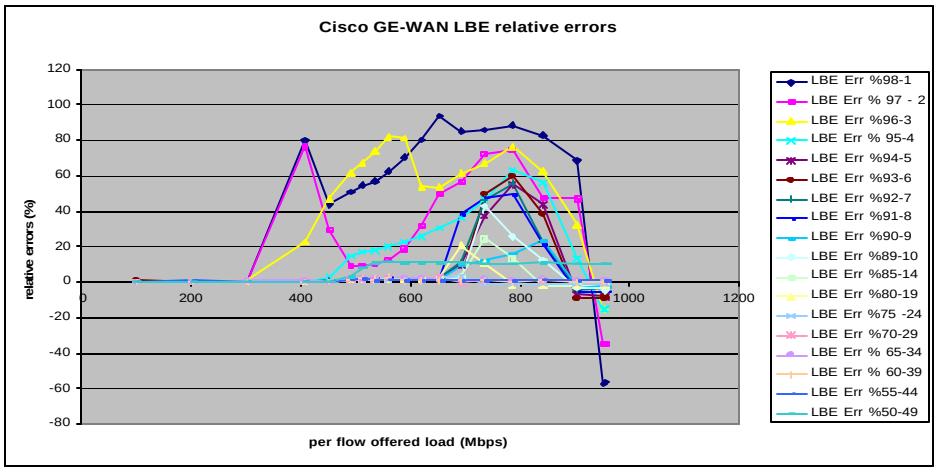

Figure 11 : The per-bandwidth allocation couple relative errors against the port congestion level for LBE

11 respectively in order to determine whether the errors were localized, whether in one or more port congestion level zones and if the error decreased monotonically with the increase of the bandwidth allocated to LBE and for a fixed value of the port congestion level.

The figures clearly show how the BE relative error is negligible (<2.5%) while that of LBE is not. The latter do not even show a monotone decrease of the error per bandwidth allocation couple and per port congestion level. The MAX LBE relative error with sign figure (Figure 12) is therefore necessary to work the boundary of the operating region.

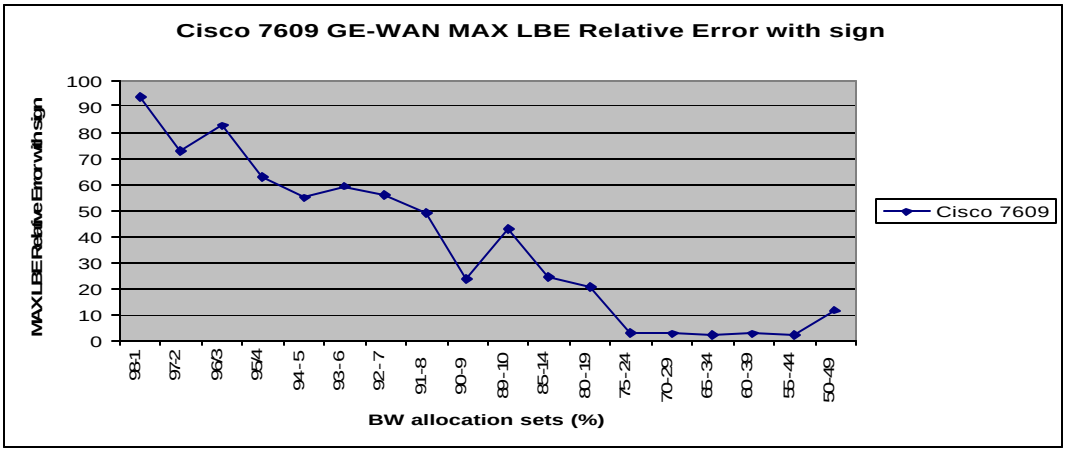

Figure 12 : The MAX LBE relative error with sign (M-LREWS) for the GigE line card

The “max” operating region for this card can be seen to be from the value of 55-44 to that of 70-29 for the bandwidth allocation couples.

3.7.2 Juniper

3.7.2.1 Testbed

Figure 13 shows the layout of the test-bed layout for the Juniper OC-48 card tests.

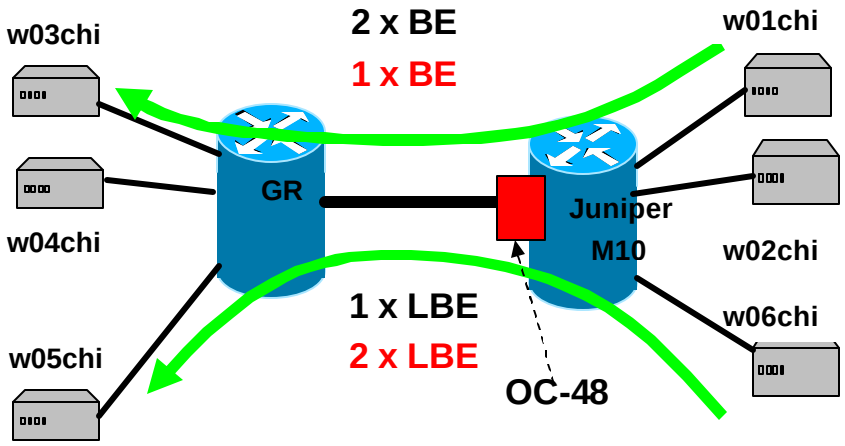

Figure 13 : The layout of the test-bed for the Juniper OC-48 line card tests

The router operating system used was "Junos 5.3R2.4" and the line card version was "1xSTM-16 SDH, SMSR-REV 05"

Figure 14 shows the layout of the test-bed layout for the Juniper GigE line card tests.

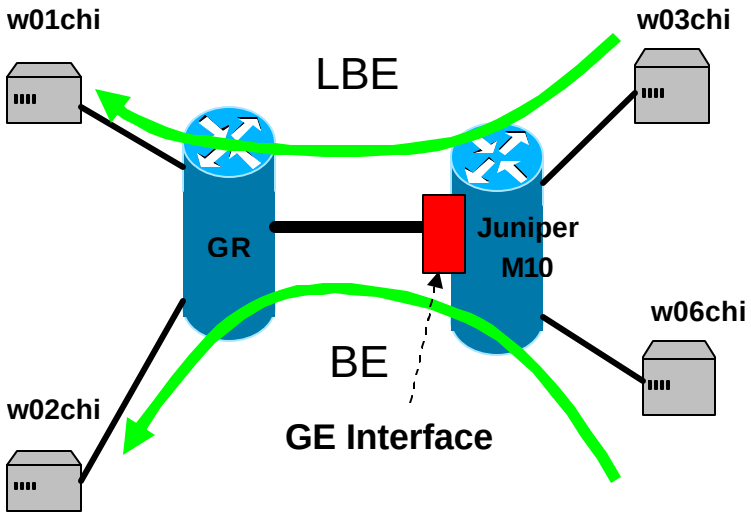

Figure 14 : The layout of the test-bed for the Juniper GigE line card tests

The router operating system used was the same as for the Juniper OC-48 test while the card version was 1x G/E, 1000 BASE-SX REV 01

3.7.2.2 OC-48 and 1GigE Configuration

The recommended configurations used for the OC-48 and for the GigE-WAN was as follows,

```
class-of-service {
    classifiers {
        dscp UCL-classifier {
            forwarding-class LBE {
                loss-priority low code-points cs1;
            }
            forwarding-class best-effort {
                loss-priority low code-points 000000;
            }
        }
    }
    forwarding-classes {
        queue 2 LBE;
        queue 0 best-effort;
    }
    interfaces {

        input {
            unit 0 {
                classifiers {
                    dscp UCL-classifier;
                }
            }
        }
        output {
            scheduler-map MAP-UCL;
            unit 0 {
                classifiers {
                    dscp UCL-classifier;
                }
            }
        }

    scheduler-maps {
        MAP-UCL {
            forwarding-class LBE scheduler sch-LBE;
            forwarding-class best-effort scheduler sch-BE;
        }
    }
    schedulers {
```

```
            sch-BE {
                transmit-rate percent X;
                buffer-size percent X;
                priority high;
            }
            sch-LBE {
                transmit-rate percent Y;
                buffer-size percent Y;
                priority low;
            }
        }
    }
```

[Note : Juniper routers have a priority queuing mechanism which is not a strict priority mechanism. The queue weight ensures the queue is provided a given minimum amount of bandwidth which is proportional to the weight. As long as this minimum has not been served, the queue is said to have a "positive credit". Once this minimum amount is reached, the queue has a "negative credit". A queue can have either a "high" or a "low" priority. A queue having a "high" priority will be served before any queue having a "low" priority.

For each packet, the WRR algorithm strictly follows this queue service order:

1. High priority, positive credit queues;
2. Low priority, positive credit queues;
3. High priority, negative credit queues;
4. Low priority, negative credit queues.

The following explanation tries to clarify the WRR mechanism.

The positive credit ensures that a given queue is provided a minimum bandwidth according to the configured weight (for both high and low priority queue). On the other hand, negative credit queues are served only if one positive credit queue has not used its whole dedicatedbandwidth and no more packets are present in a "positive credited" queue.

The leftover bandwidth (from the positive credited queues) is fairly shared between all the "high priority negative credit" queues until these ones become empty. If the high priority negative credit queues are empty and if there is still some available bandwidth that can be allocated to packets, the "low priority negative credit" queues will equally share it.

The credits are decreased immediately when a packet is sent. They are increased frequently.

The last thing to mention is that the "maximum-buffer-delay percent x" command does NOT associate a buffer length to a queue. RED has to be used if such association has to be enforced.

It is worth noticing that the best QoS configuration expects low priority ("priority low") for the class which is allocated less bandwidth and vice versa high priority ("priority high") to the class which is allocated more bandwidth. This is necessary for Juniper in order to precisely allocate a class (BE in our case) the minimum guaranteed bandwidth even when this is very small (<5%).

This is something which only applies to the way Juniper implements the scheduler and it is not therefore a reasoning of general validity.

It is important to say that the "side-effect" of this command line entry would also be that of entirely assigning any non-allocated (see "spare" in the "Error Analysis paragraph) minimum bandwidth to the higher priority class when the interface is congested and both classes are over-subscribed. But this situation didn't actually happened during the test as no spare capacity ("spare"= 0) was left since Juniper is the only manufacturer out of the three of them that makes possible to reserve up to 100% of the port capacity.]

3.7.2.3 OC-48 Results

Figures 15 and 16 show link utilisation as a function of per-flow offered load for a range of BE/LBE allocations for (2BE+1LBE) and (1BE+2LBE) flows respectively

The iperf UDP-payload-level capacity of the link, which was obtained by congesting the interface and not configuring qos was 2338 Mbps. The card is therefore congested up to (957*3)/2338 = 2871/2338 = 122% of its capacity.

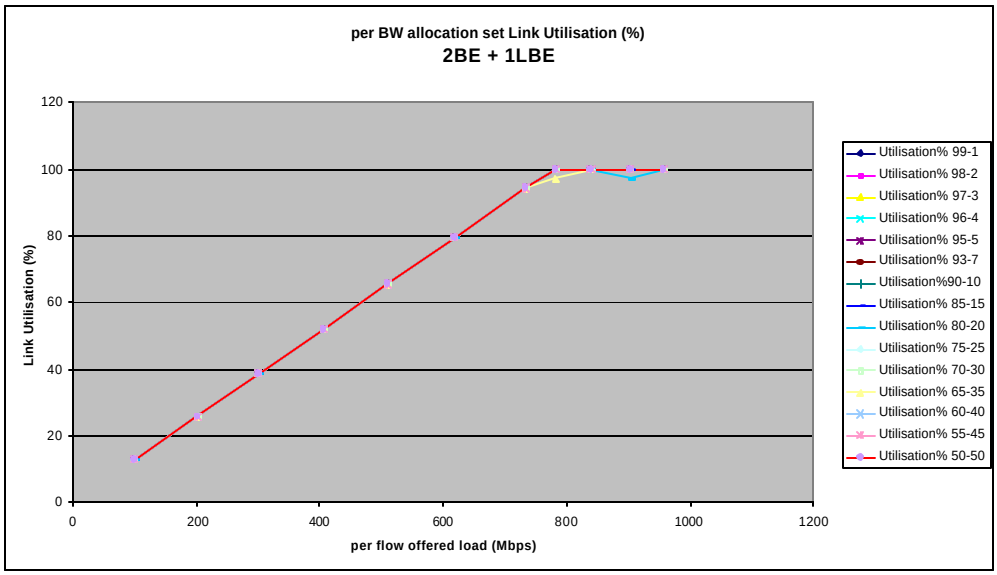

Figure 15 : Link utilisation as a function of per-flow offered load for a range of BE/LBE allocations for (2BE+1LBE) flows

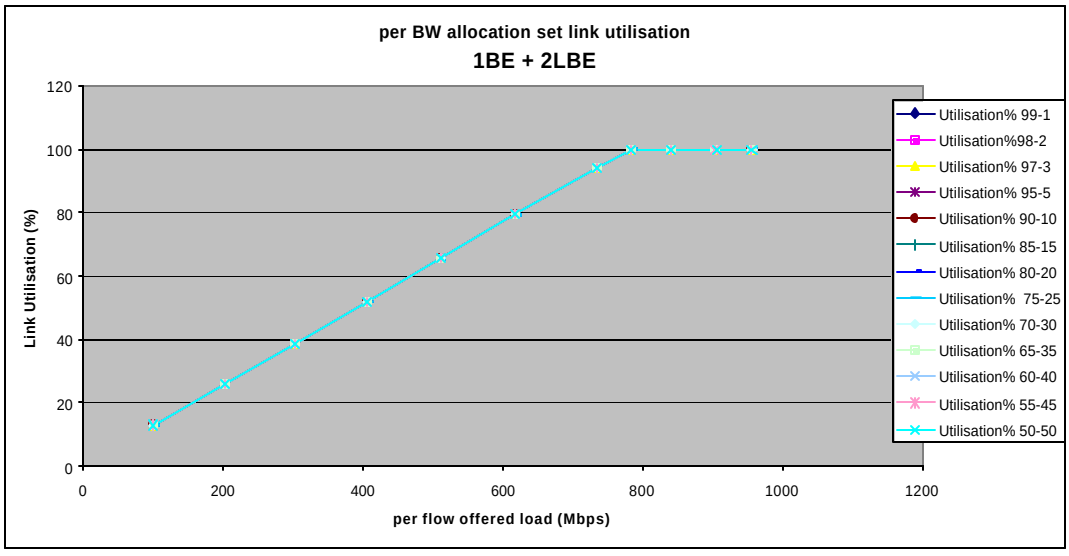

Figure 16 : Link utilisation as a function of per-flow offered load for a range of BE/LBE allocationsfor (1BE+2LBE) flows

As a good link utilisation is not sufficient to have a good bandwidth allocation precision, the per-bandwidth allocation couple relative errors for both BE and LBE are presented in Figures 17 and 18 against different levels of port congestion.

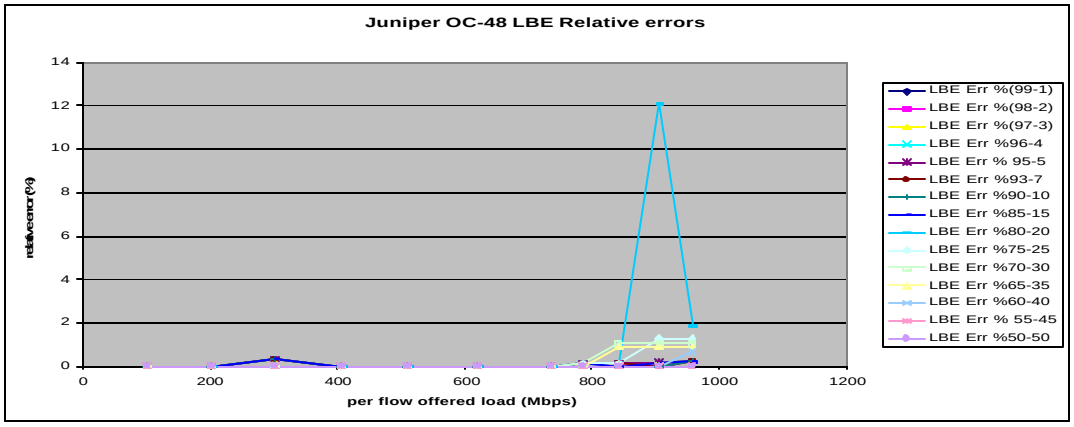

Figure 17 : The per-bandwidth allocation couple relative errors against the port congestion level for LBE

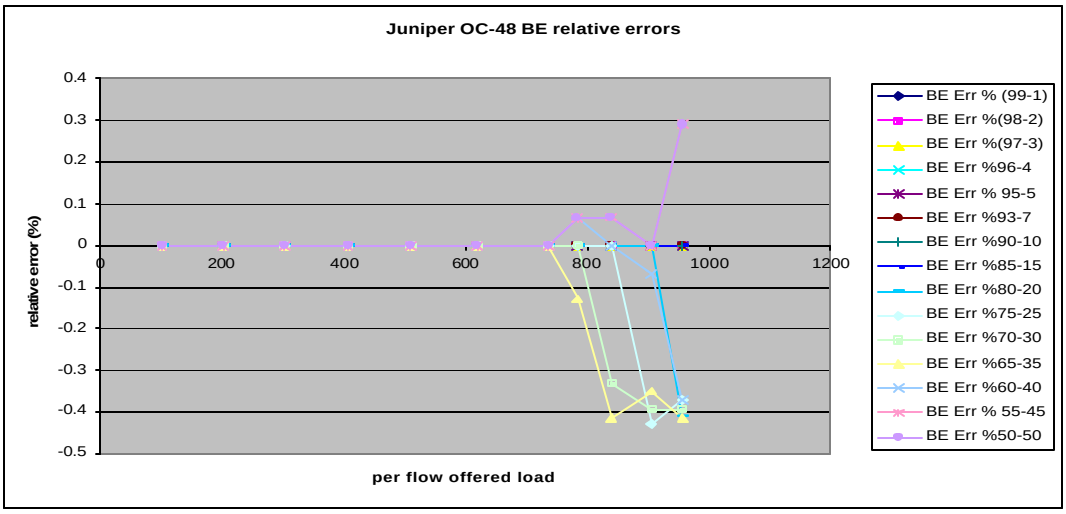

Figure 18 : The per-bandwidth allocation couple relative errors against the port congestion level for BE

Setting aside some problems associated with poor background data traffic performances issues, both BE and LBE error is negligible. Therefore the whole bandwidth allocation set is a “max” operating region and Figure 19 presenting the MAX LBE Relative error with sign confirms this finding.

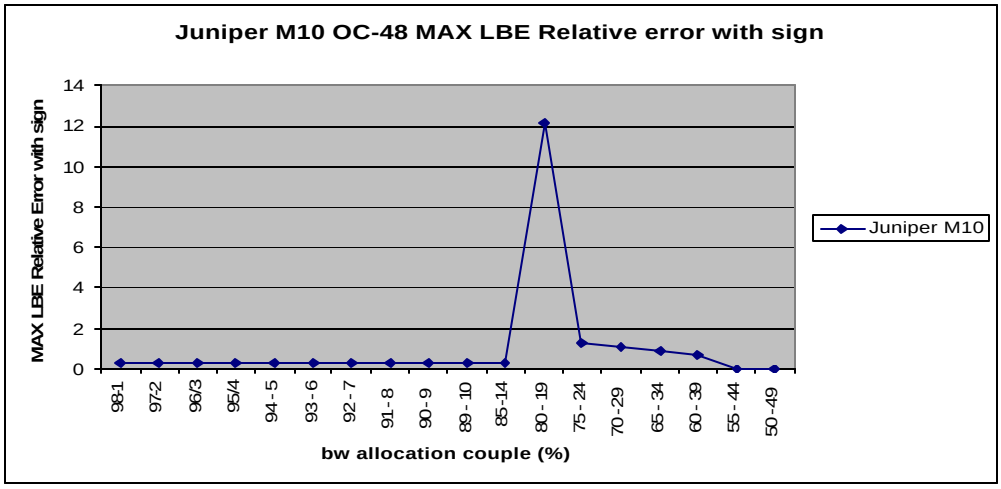

Figure 19 : The MAX LBE Relative error with sign for the Juniper OC-48 line card

3.7.2.4 1GigE Results

Figure 20 shows the layout of the testbed for the Juniper GigE line card tests.

The iperf UDP-payload-level capacity of the link, which was obtained by congesting the interface and not configuring QoS was 957 Mbps. The card was therefore congested up to (957*2)/957=200% of its capacity.

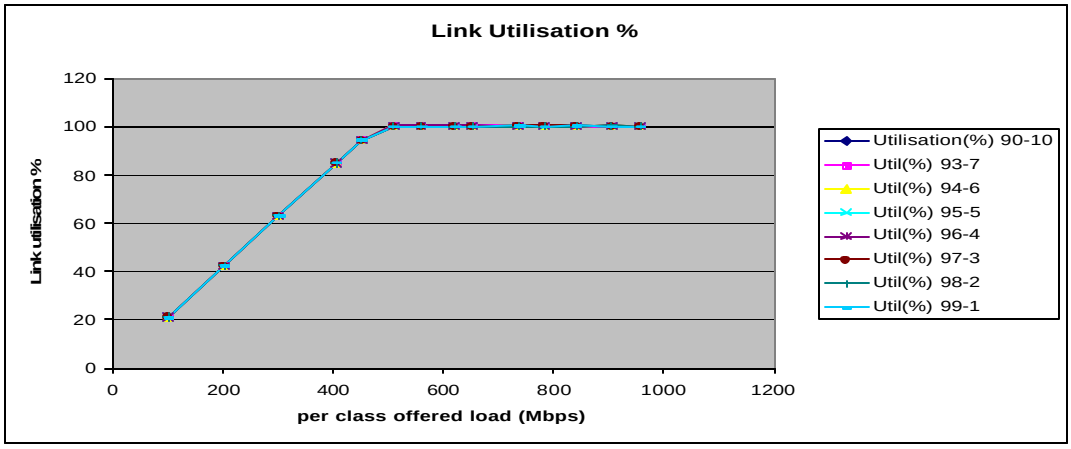

Figure 20 : The layout of the testbed for the Juniper GigE line card tests.

Again the relative BE and LBE bandwidth allocation precision errors need to be reviewed in order to see whether there are errors and thus their possible magnitude and dynamics along the bandwidth allocation couples and along the port congestion levels regions. These are shown in Figures 21 and 22.

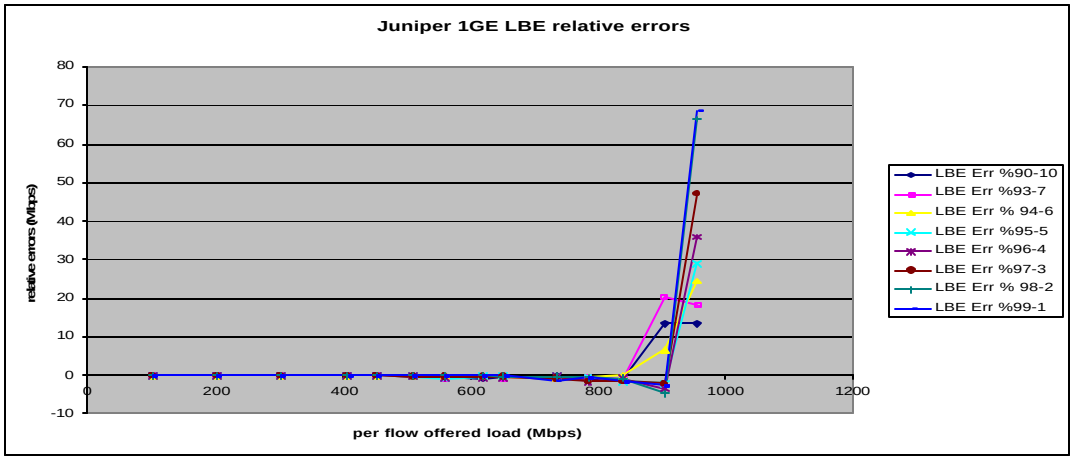

Figure 21 : The relative LBE bandwidth allocation precision errors for the GigE Line card

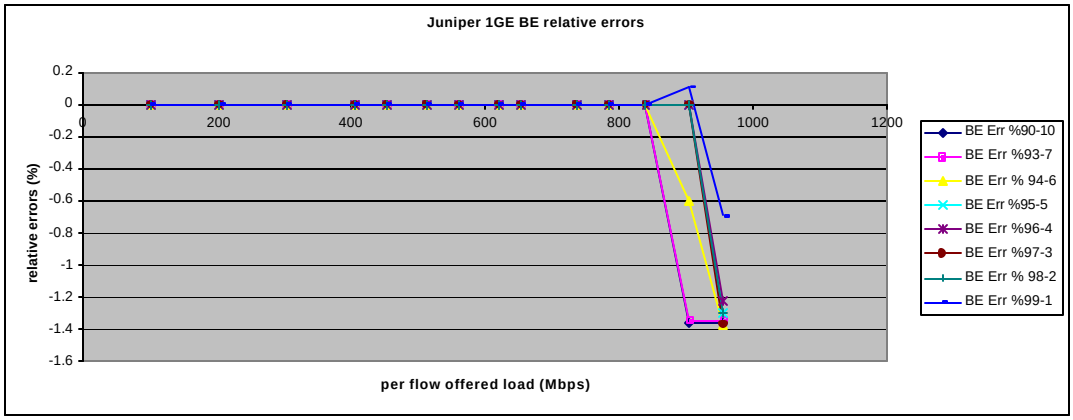

Figure 21 : The relative LBE bandwidth allocation precision errors for the GigE Line card

It is clear from these Figures that the BE error is negligible and mainly negative while the LBE error is mainly positive and is not negligible. The latter error decreases monotonically with the increase of the bandwidth allocation couples, this suggesting that the MAX LBE Relative error is monotone as well which is confirmed from Figure 22 which shows that the "interpolated" "max" operating region of the bandwidth allocation couples ranges from 50-49 to 70-29.

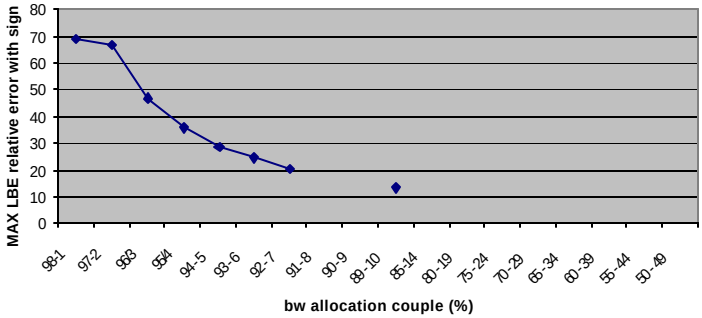

Figure 22 : The MAX LBE Relative error of the Juniper GigE line card

3.7.3 Procket

3.7.3.1 Testbed

Figure 23 shows the layout of the testbed for both Procket OC-48 and 1G igE line card tests.

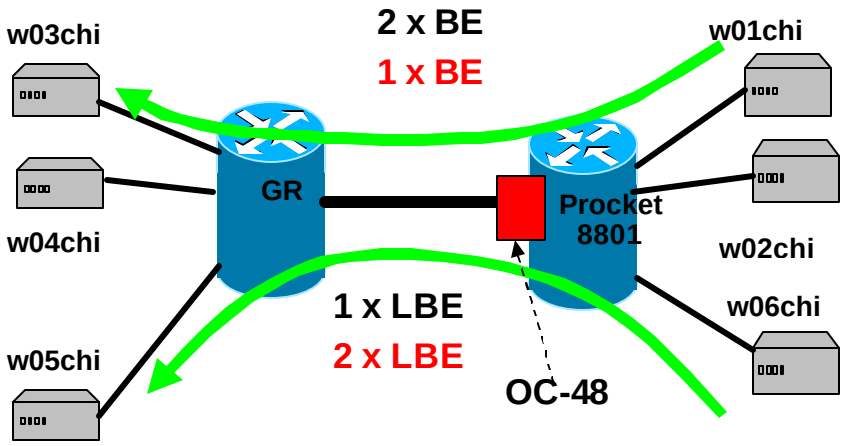

Figure 23 : The layout of the testbed for both Procket OC-48 and 1GigE line card tests

The System Release Version used was 2.3.0.180-B and the Kernel Version used was 2.3.0.1-P PowerPC while the line card versions were 4-PORT OC-48c POS SR and 10-PORT 1000BASE-SX for the 1GigE and OC-48 port respectively.

3.7.3.2 OC-48 and 1GigE Configuration

The recommended configurations used for the OC-48 and for the GigE-WAN was as follows,

```
!
qos
 class BE
  dscp 0
 class LBE
  dscp 8
 service-profile UCL
  class BE
  class LBE
  queuing-discipline dwrr (BE[X], LBE[Y], default[1])

!
interface output
qos-service UCL
!
```

3.7.3.3 OC-48 Results

Figure 24 shows the link utilisation as a function of the per-flow offered load for a range of BE/LBE flows

The iperf UDP-payload-level capacity of the link, which was obtained by congesting the interface and not configuring QoS was 2337 Mbps. The card was therefore congested up to (957*3)/2337 = 2871/2337 = 122% of its capacity.

Figures 25 and 26 show the BE and LBE Relative errors with sign respectively for the Procket OC-48 line card and allow the assessment of errors in the precision of the bandwidth allocation are present.

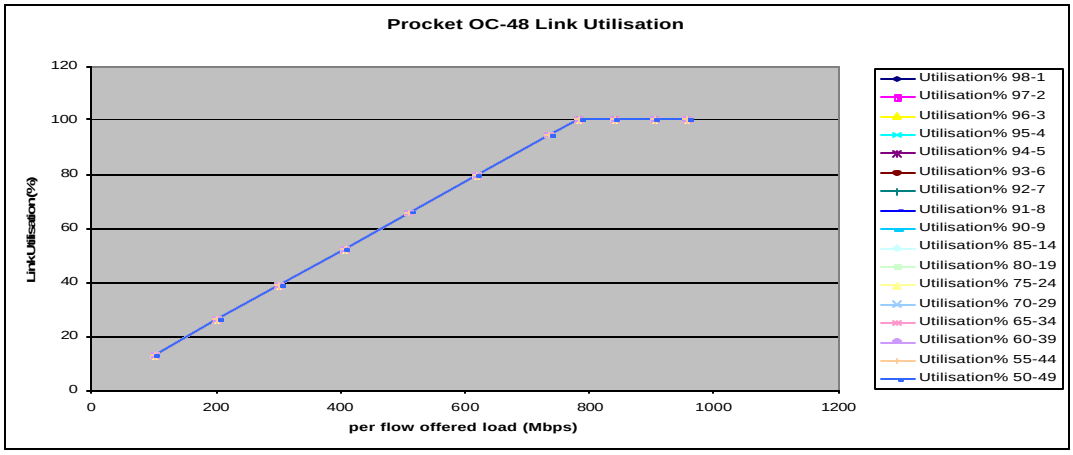

Figure 24 : The link utilisation as a function of the per-flow offered load for a range of BE/LBE flows

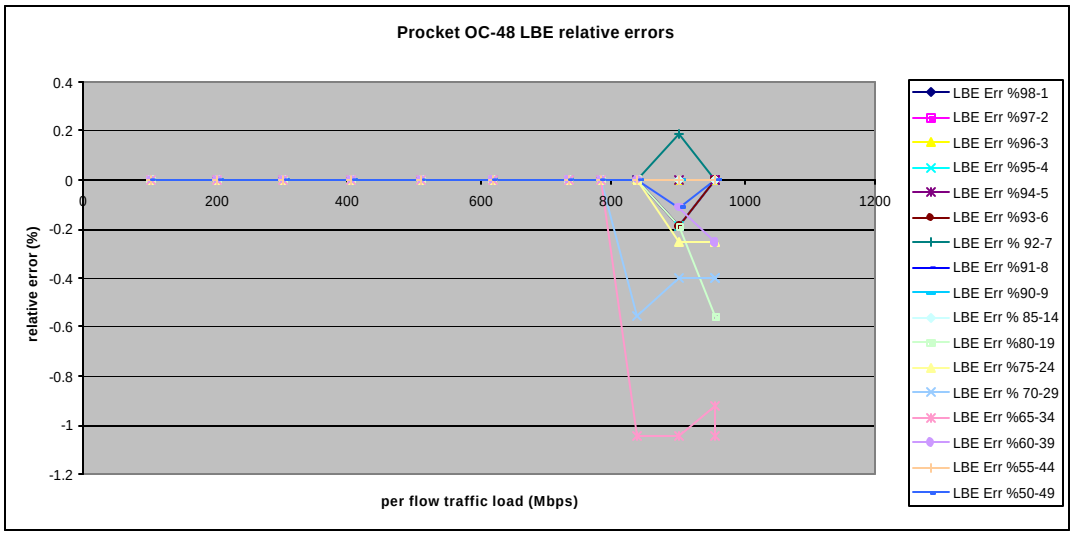

Figure 25: The LBE relative errors as a function of per-flow traffic loads for the Procket OC-48 line card

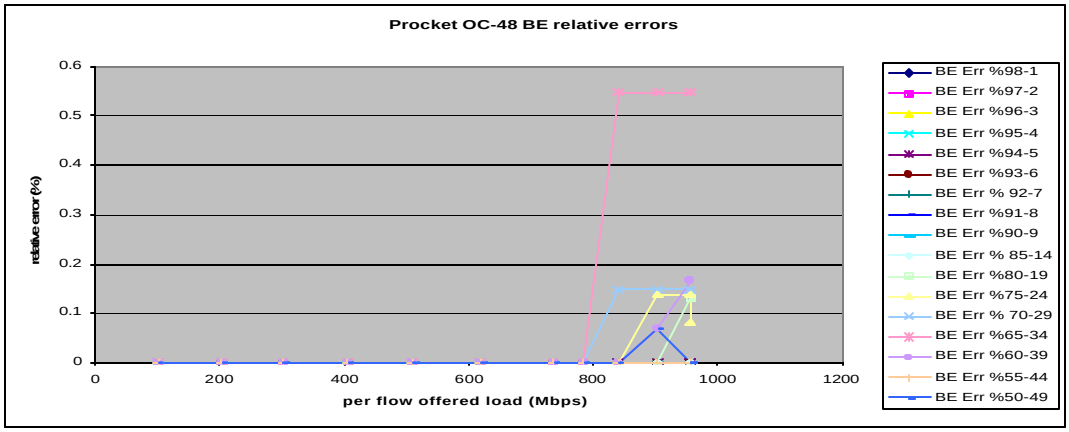

Figure 26: The BE relative errors as a function of per-flow traffic loads for the Procket OC-48 line card

The link utilization is perfect and both BE and LBE show negligible errors (<1%). The interesting thing is that such errors appear from 80-19 towards 50-49 for both classes and that BE is actually positive while LBE is negative. The exact opposite error polarization if compared with the typical errors the other manufacturers show.

3.7.3.4 1GigE Results

Figure 27 shows the link utilisation as a function of the per-flow traffic load for a range of BE/LBE allocations.

The iperf UDP-payload-level capacity of the link, which was obtained by congesting the interface and not configuring QoS was 957 Mbps. The card was therefore congested up to (957*3)/957=300% of its capacity. It is worth noting that this card was congested up to 300% (test1) of its capacity which was 100% more congested than the maximum congestion experienced by both GigE Juniper and GigE-WAN Cisco.

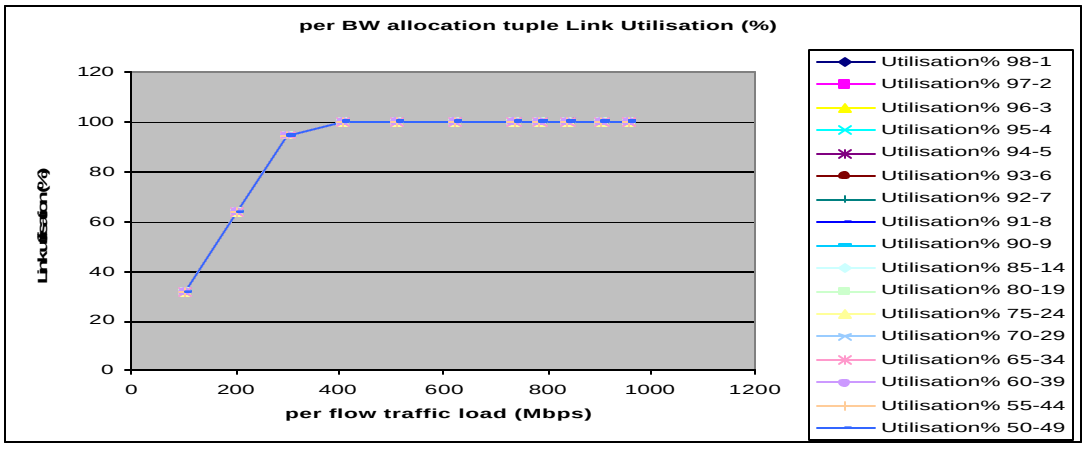

Figure 27 : T he link utilisation as a function of the per-flow traffic load for a range of BE/LBE allocations.

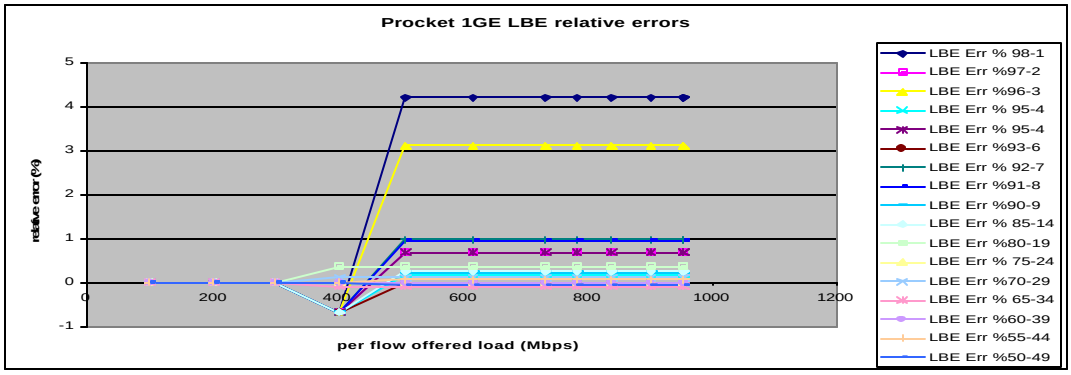

Figure 28 : The LBE relative errors as a function of the per-flow offered load for the Procket 1GigE line card

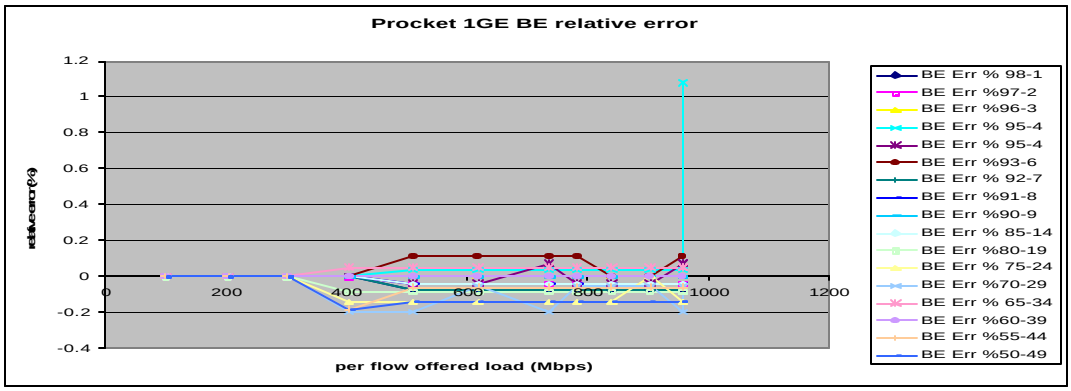

Figure 29 : The BE relative errors as a function of the per-flow offered load for the Procket 1GigE line card

The Link Utilisation is once again perfect, the BE relative errors are negligible and the LBE ones rapidly tend to become negligible. The MAX LBE relative error with sign (M-LREWS) plotted against the bandwidth allocation couples is presented in Figure 30 and aside from 98-1 and 96-3 all other couples show an error of less than 1%. The operating region ranges therefore from 95-4 to 50-49 out of the whole bandwidth allocation couples axis.

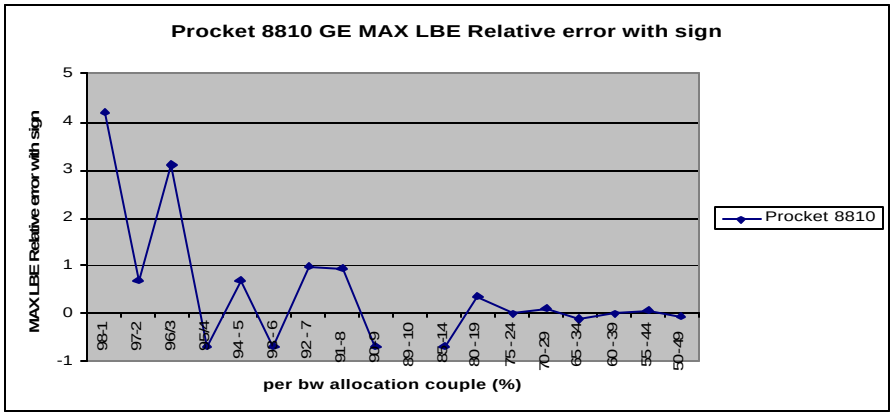

Figure 30 : The relative error calculated as a function of per-bandwidth allocation for the Procket GigE line card

3.8 COMPARATIVE ANALYSIS

From the foregoing analysis, it can be seen that the majority of the errors are localised on LBE and therefore the relative error will be used to compare the performances of the different routers.

It is worth noting that most of the manufacturers fix the accuracy in allocating bandwidth within a class to be around 95% which in turn means that 2.5% is the maximum error acceptable in a module. The bandwidth allocation couple region inside which this bound is validated is referred to here as the "operating region".

In order to bound the operating region out of the bandwidth allocation couples axis, the M-LREWS (Max LBE Relative Error With Sign) metric is used for both GigE and OC-48 and for all the three router manufacturer involved. This allows the evaluation of the bandwidth scheduler based solely on its worst performance over the port congestion level axis. This is of extreme importance since the bounded value for the precision in the allocation of bandwidth that the manufacturers refer to can be correctly associated to the worst case scenario out of the whole offered load axis. It is therefore correct to say that an accuracy of 95% in the allocation of the bandwidth is equivalent to have the M-LREWS/M-BREWS <= +-2.5%. The so defined operating region is referred to here as the "max" operating region, thus highlighting that the method used to bound it was that of computing the max algebra for the relative errors.

In order, then, to evaluate the performances of a line card averaged over the whole offered load axis, or better, averaged over different line card congestion levels, the AALRE (Average Absolute LBE Relative Error) metric is presented for both GigE and OC-48 and for all the three router manufacturers under test.

The absolute values are used to avoid the situation when average of algebraic values could lead to a misleading ~0% error as discussed above. The so defined operating region is referred to here as the "avg" operating region, this highlighting that the method used to bound it was that of computing the avg algebra for the relative errors.

3.8.1 M-LREWS (Max LBE Relative Error With Sign)

3.8.1.1 OC-48 M-LREWS

Figure 31 shows for OC-48 line cards of each router the maximum LBE relative error with sign as a function of the bandwidth allocation.

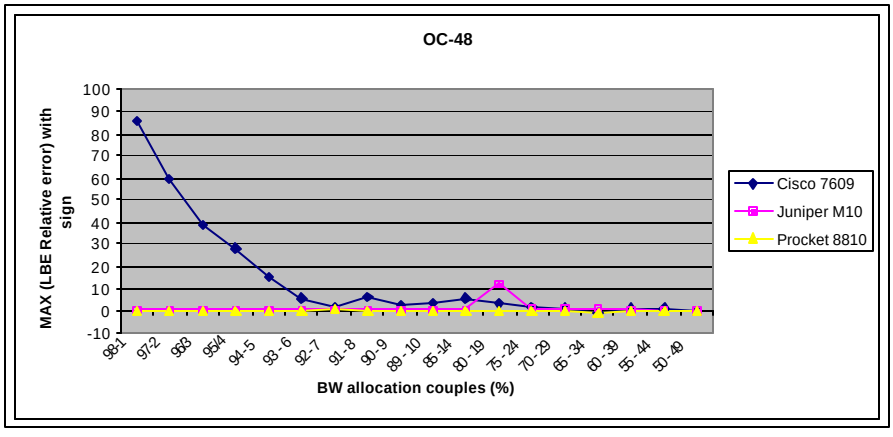

Figure 31 : For OC-48 line cards of each router the maximum LBE relative error with sign as a function of the bandwidth allocation.

In order to work out which operating region applies to the different manufacturers, a zoom over the abscissa region where all the three curves are close to the value of 2.5 is shown in Figure 32.

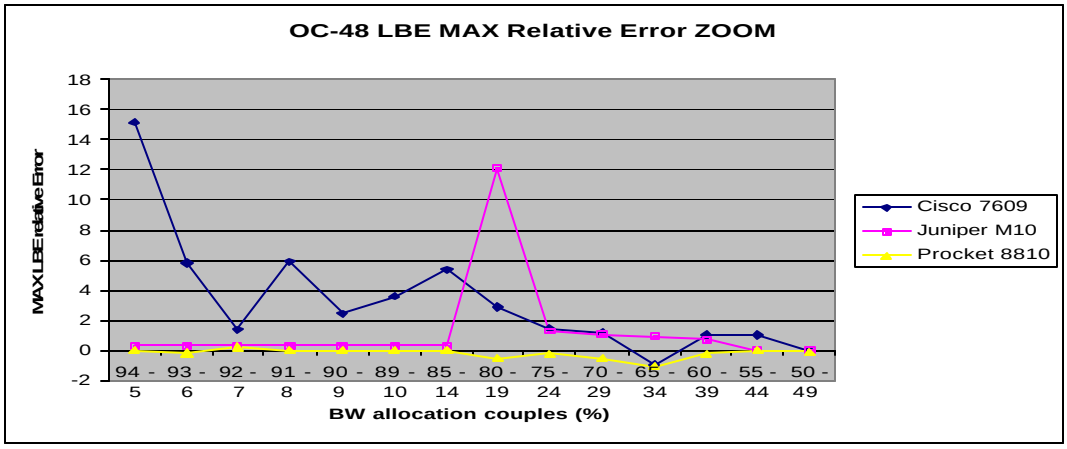

Figure 32 : A more detail view of the abscissa region from Figure 31.

Apart from a glitch showed by Juniper, the entire bandwidth allocation couplet axis is a "max" operating region for both Juniper and Procket with the latter performing slightly better. It is difficult to determine a "max" operating region for the Cisco as the error does not decrease monotonically oscillates around the value 2.5. As a consequence, a conservative "max" operating region over the bandwidth allocation couple axis is that which ranges from 75-24 to 50-49.

3.8.1.2 GigE M-LREWS

Figure 33 shows for GigE line cards of each router the maximum LBE relative error with sign as a function of the bandwidth allocation. It is worth highlighting here that the Procket card was congested up to 300% of its capacity while the maximum congestion that Cisco GigE-WAN and Juniper GigE experienced during the test was only 200%.

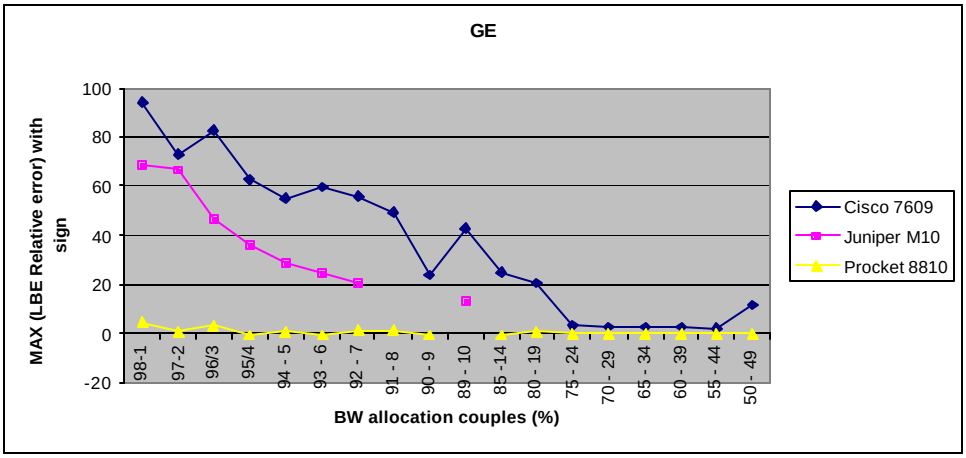

Figure 33: GigE line cards of each router the maximum LBE relative error with sign as a function of the bandwidth allocation

With the target accuracy fixed to the canonical 95%, Cisco "max" operating region, out of the whole bandwidth allocation couples axis, ranges from 70-29 to 55-44. Juniper "max" operating region, which is linearly interpolated out of the values obtained, ranges from 70-29 to 50-49 although its performance is better than the Cisco throughout most of the bandwidth allocation couples axis. Procket "max" operating region ranges from 95-4 included to 50-49.

In order to take into account the overall performance for different and increasing port congestion levels which may change the results obtained with the max analysis, an average (AVG) analysis follows.

3.8.2 A-ALRE (Average Absolute LBE Relative Error)

3.8.2.1 OC-48 A-ALRE

Figure 34 shows the average absolute LBE relative error as a function of the bandwidth allocation for the three routers tested for the OC-48 line cards.

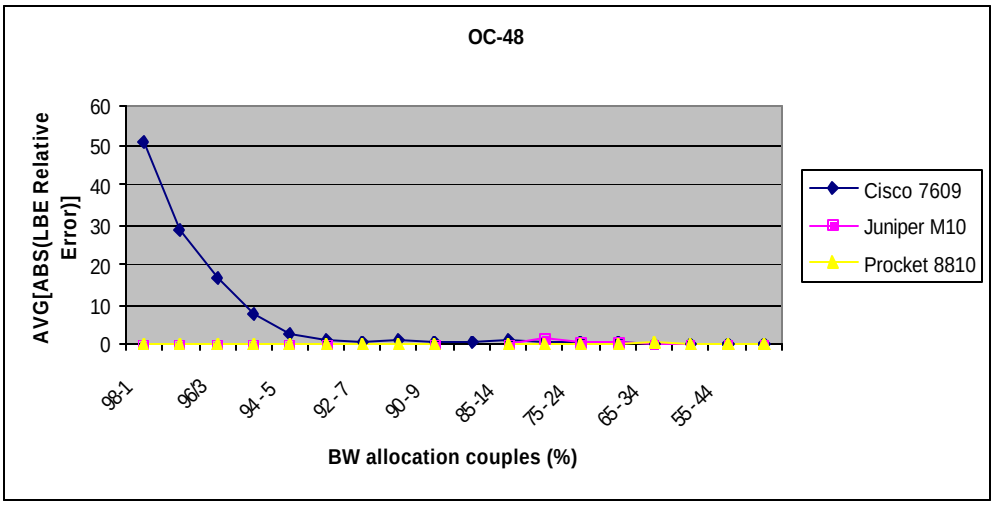

Figure 34 : The average absolute LBE relative error as a function of the bandwidth allocation for the three routers tested for the OC-48 line cards.

Figure 34 shows that the Cisco operating region averaged over the whole port congestion levels axis ("avg" operating region) ranges from 94 – 5 included to 50-49. It is worth noticing how the average lowers the values but also acts, in this case, as a low pass filter whose effect is that of smoothing out the oscillations that led before to a conservative evaluation of the Cisco "max" operating region and that was the main reason for such a poor performance evaluation. The Cisco "avg" operating region is, in fact, much better than the "max" operating region which spanned from 75-24 to 50-49.

Figure 35 zooms in on the lower two curves of Figure 34 and shows how the error is negligible for both although the Procket router shows again slightly better performance.

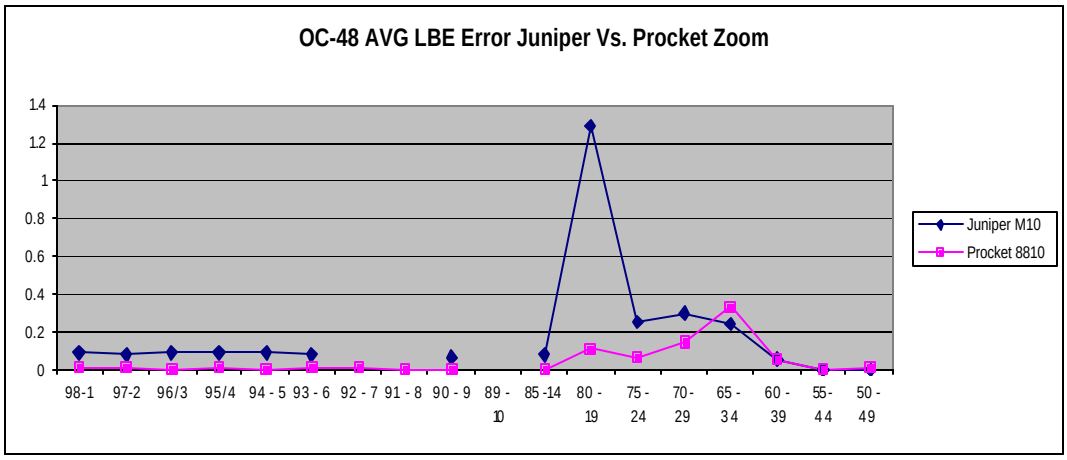

Figure 35 : A more detail view of the abscissa region from Figure 34.

3.8.2.2 GigE A-ALRE

Figure 36 shows the average absolute LBE relative error as a function of the bandwidth allocation for the three routers tested for the GigE line cards.

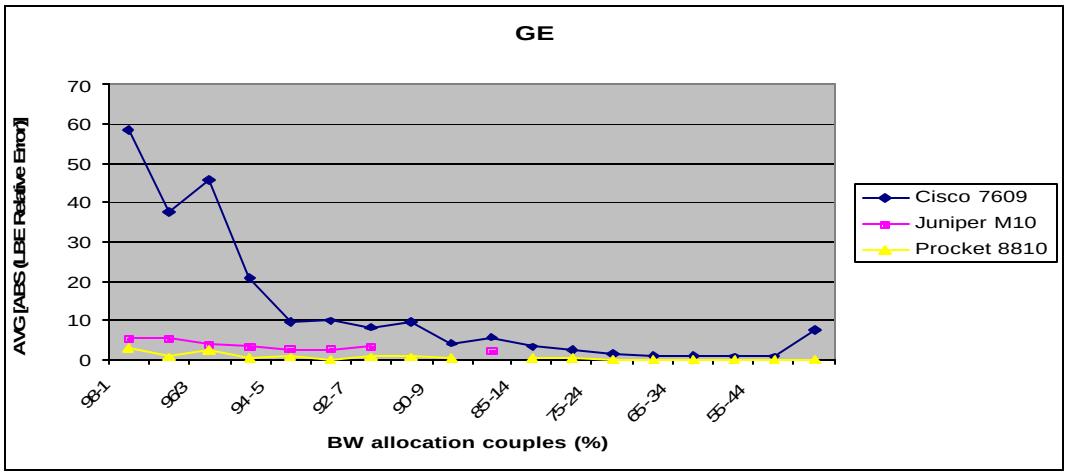

Figure 36 : The average absolute LBE relative error as a function of the bandwidth allocation for the three routers tested for the GigE line cards.

It is worth noticing that, again, the average performance of both Cisco GigE-WAN and Juniper M10 GigE are much better than their relative "max" performance proving that the error is not spread along the offered_load/port_congestion_levels axis. Cisco "average" operating region ranges from 75-24 included to 55-44 which is 26% of the bandwidth allocation couple axis. In order to work out the "avg" operating region for both Juniper and Procket, a zoom is needed and is presented in Figure 37.

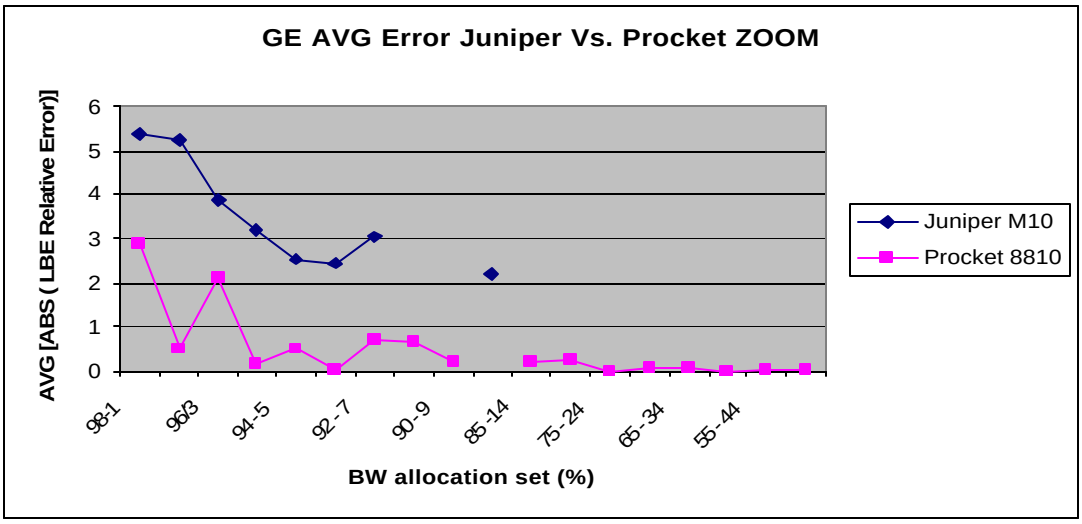

Figure 37 : A more detail view of the abscissa region from Figure 36.

The Procket "average" operating region ranges from 97-2 included to 50-49 while the Juniper interpolated "average" operating region ranges from 91-8 included to 50-49.

It is worth highlighting how the overall OC-48 line cards performance is better than that of the GigE line cards for both the "MAX" and "AVG" algebra analysis. This suggests that it is not raw speed that is the dominant issue in order to design a good bandwidth IP-level scheduler but that it is the link protocol underneath. In particular, OC-48 presents SONET as the link protocol where the two main differences between the two link technologies being that SONET employs a serial synchronous transmission while GigE employs an asynchronous serial transmission and that SONET/SDH is a much mature technology to operate and scale at Gbps speed rather than as the point to point Ethernet.

3.9 CONCLUSIONS

Both OC-48 and GigE cards from each router manufacturer have been benchmarked by looking at the achieved link utilisation and at how the BE and LBE Relative errors change over the bandwidth allocation set axis with an increasing level of port congestion.

This study has highlighted how a good link utilisation is necessary but is not the only determinant to a precise bandwidth allocation. The study suggests that the error dynamic per bandwidth allocation couple and per-port congestion level are necessary indicators in order to evaluate if and where errors in the allocation of the minimum guaranteed bandwidth under congestion occur.

The evaluation of the line card performance based on an accuracy in the allocation of the bandwidth of 95% was chosen. This is equivalent to have the maximum LBE relative error with sign (M-LREWS) < +- 2.5%; and for this reason it is called "max" operating region (over the bandwidth allocation couples).

BE error is not taken into account as it is always almost negligible for any of the line cards under test. This result suggests that the main problem these cards encountered was their inability to re-allocate the left-over BE bandwidth to LBE. This resulted in a narrower operating region available as a consequence. The following table (Table 1)summarises these results for both OC-48 and GigE line cards

OC-48	Cisco	Juniper	Procket
"Max" operating region	75-25 to 50-50. 6/19=31.5%	99-1 to 50-50 100%	99-1 to 50-50 100%

GigE	Cisco	Juniper	Procket
"Max" operating region	70-30 to 55-45 4/19=21%	70-30 to 50-50 5/19=26.3%	95-5 to 50-50 14/19=73.6%

Table 1 Results summary for the performance of both OC-48 and GigE line cards for Cisco, Juniper and Procket routers (see test for detail)

The results show that Procket has the best performances for both line cards and with the OC-48 line card the results were perfect.

The results for the Juniper were very close in performances to Procket for the OC-48 line card but very close to Cisco for the GigE line card.

The results for Cisco were the worst performances of the three manufacturers for both line cards.

The A-ALRE analysis was then considered in which the LBE relative error is averaged over the whole port congestion level axis. This leads to a comparison based on both errors. Table 2 shows the relative table along with the computation of the percentage improvement (delta ?) in passing from the "max" to the "avg" operating region for both OC-48 and GigE line cards.

OC-48	Cisco	Juniper	Procket
Max op region	75-25 to 50-50. 6/19=31.5%	99-1 to 50-50 100%	99-1 to 50-50 100%
Avg op region	94 – 6 to 50-50 13/19=68.42% ?=+117.2%	//	//

GigE	Cisco	Juniper	Procket
Max op region	70-30 to 55-45 4/19=21%	70-30 to 50-50 5/19=26.3%	95-5 to 50-50 14/19=73.6%
Avg op region	75-25 to 55-45 5/19=26.3% ?=+6%	91-9 to 50-50 10/19=52.6% ?=+100%	97-3 to 50-50 15/19=78% ?=+6%

Table 2 : The relative values along with the computation of the percentage improvement (delta ?) in passing from the "max" to the "avg" operating region for both OC-48 and GigE line cards.

What is of particular interest is that the improvement delta of 100% for Juniper in passing from the "max" to the "avg" operating region in comparison to the 6% delta improvement achieved for Cisco

for the same passage. This suggests that the Cisco LBE relative error is much more widely spread and therefore serious all over the entire port congestion level axis in comparison with that of Juniper which is much more localised across fewer port congestion levels.

It is also clear, for the three manufacturers, that the QoS implementation in the OC-48 line cards presents a much more precise formulation than that found for the GigE line cards. This suggests that raw speed is not the main issue in the design of good bandwidth schedulers but that the link layer technology which underlies the IP-layer bandwidth scheduler has more relevance.

It is however true that for the tests of the GigE line cards the level of over-commitment was greater than for the equivalent OC-48 line card tests, i.e. 3 * 1Gpbs over a 1Gbps link as opposed to 3 * 1Gbps over a 2.5Gbps link. This may be of significance but the test environment was the same for all line cards tested.

The fact that SONET employs a synchronous serial transmission while GigE uses an asynchronous serial transmission may also be of significance to these results

Finally, SONET is a much more mature technology operating at Gigabit rates in comparison with GigE and this may contribute in some way to the results presented here.

3.10 FUTURE WORK - SYNTHESIS

The work described here is significant as it provides the essential background for the deployment of QoS in a network with respect to the performance of different line cards from a range of router vendors. For the Grid, where end-to-end network performance matters and, where increasingly the network will include high bandwidth components often scaled across great distance, QoS is but one component of performance. DataTAG WP2 has also been involved in the investigation of developments to TCP to operate within the high bandwidth, high RTT environments where standard TCP operates ineffectively.

Work is now underway to investigate the capabilities of QoS using the new TCP stacks both individually and as one component in more generalised TCP traffic. This synthesis of the work of DataTAG WP2 will be crucial for the deployment of solutions that meet the needs of Grid applications in, for example, particle physics, astronomy and biology.

In more detail, the investigations will explore the relationship between IP-QoS configuration in the routers and the dynamics of new proposals for high throughput TCP, including High Speed TCP, Scalable TCP and FAST. These tests will make use of the DataTAG testbed using the Juniper M10 routers with differentiated services enabled GigE line cards (the choice being made following the line card benchmarking described here).

Three IP QoS classes have been configured: BE for traditional Best Effort traffic (WEB-like traffic), AF for TCP flows associated with the new networked GRID applications and EF for real-time applications. By varying the rates of the flows in these classes, it has been possible to measure how responsive the specific TCP stack is to changes in the available capacity.

The results obtained so far suggest a mechanism for segregating traffic sharing the same packet-switched path is needed. This is required for two reasons. Firstly to protect traditional BE traffic from the more aggressive AIMD approach taken by the new TCP stacks. Secondly to guarantee a level of end-to-end service predictability for applications making use of the new TCP stacks which is sufficient to enforce a network resources reservation system through the use of the GRID middleware.

Much work remains to be done to verify and to extend these early results and this will work will be the focus of DataTAG WP2 Task 2.2 activity between now and the EU Review in March 2004 where a detailed analysis will be available for consideration.

4 EQUIVALENT DIFFERENTIATED SERVICES

4.1 INTRODUCTION

Grid flows crossing IP networks are not equally sensitive to loss or delay variations. For several years, much research has been carried out in an attempt to solve the problem of the heterogeneous performance needs of the IP traffic. As opposed to the philosophy of over provisioning, a class of solutions have considered that the IP layer should provide more sophisticated set of services than the simple best-effort service to meet the quality of service requirements of applications. Different proposals for improving the IP stack have been proposed [DS] but these still present limitations and difficulty [Teitelbaum]. The major obstacles being that of deployment and of scaling. In the light of the experiences in deployment of existing IP QoS approaches, IntServ and DiffServ, a new differentiated service scheme called Equivalent Differentiated Services (EDS) has been developed [Hurley][Montenegro].

EDS represents a radical departure from traditional from the diffServ architecture which relies upon a bounded domain concept and associated pricing models. EDS merges and extends the Alternative Best Effort ideas [Goutelle] and the Proportional DiffServ Principles [Dovrolis] that were developed and report in Deliverable D2.1.

The EDS scheme aims to provide a spectrum of "different but equivalent" network services that offer a trade-off between delay and loss rate to the end-to-end flows. EDS acts as a network layer protocol analogous to IP such that the end-to-end transport layer has to do some adaptation. This is analogous to the operation of TCP over IP. As EDS offers a service differentiation based on packet marking, the corresponding transport layer has to adapt data transmission and packet marking accordingly. Considering that the data flows are composed of real-time traffic; interactive traffic; WEB traffic and bulk file transfer traffic, different types of adaptive packet marking algorithms, integrated in a transport protocol stub, have been design to fully exploit the differentiate behaviours of the network. An implementation based on LINUX has been developed for the DataTAG project.

This software comprises different LINUX modules:

- a novel router mechanism merging an original RED-based active queue management algorithm [RED];
- a proportional scheduling algorithm; and
- a transport protocol for bulk data transfer that integrates an adaptive packet marking algorithm in the SCTP AIMD algorithm.

The software has been functionally validated and its performance evaluated within the DataTAG project. The aim has been to validate the EDS concept and to show that it can improve the transfer of a mix heterogeneous flows over long distance, heterogeneously provisioned links with a low deployment cost.

4.2 PER HOP BEHAVIOUR - IMPLEMENTATION AND IMPROVEMENT

The EDS proposal is based on the usage of the diffServ packet marking and differentiated forwarding principles, but excludes other diffserv concepts including domain and edge admission control. The network layer of the EDS architecture has been designed, evaluated and reported upon previously.

The EDS scheduler ensures that the class of services obtain a performance (in terms of delay and loss rate) proportional to the performance obtained by another class, according to a specific coefficient. Moreover there is an asymmetry between delay and loss rate performance such that the class which obtains the i-th best performance in delay obtains the i-th worst performance with respect to loss rate. This mechanism which effectively integrates a proportional scheduler and an active queuing mechanism has been implemented and evaluated in a LINUX environment. In addition an

improvement has been made with the inclusion of a RED mechanism. This has resulted in a smoother evolution of the queue length by performing early drops in the queue.

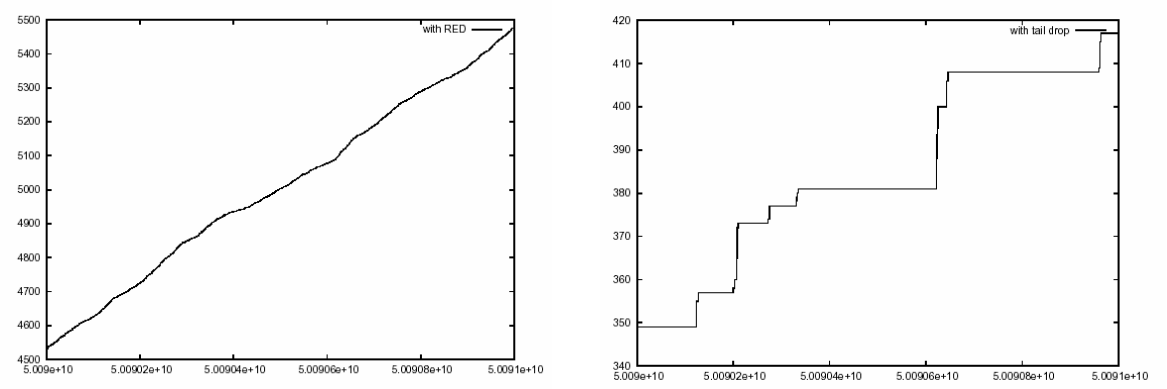

Figure 38: Losses due to RED (left hand) and losses due to trail drop (right hand)

The comparison of the loss due to RED and tail drop is shown in Figure 38 where it can be seen that the effect of RED was to smooth the loss rate in the queue.

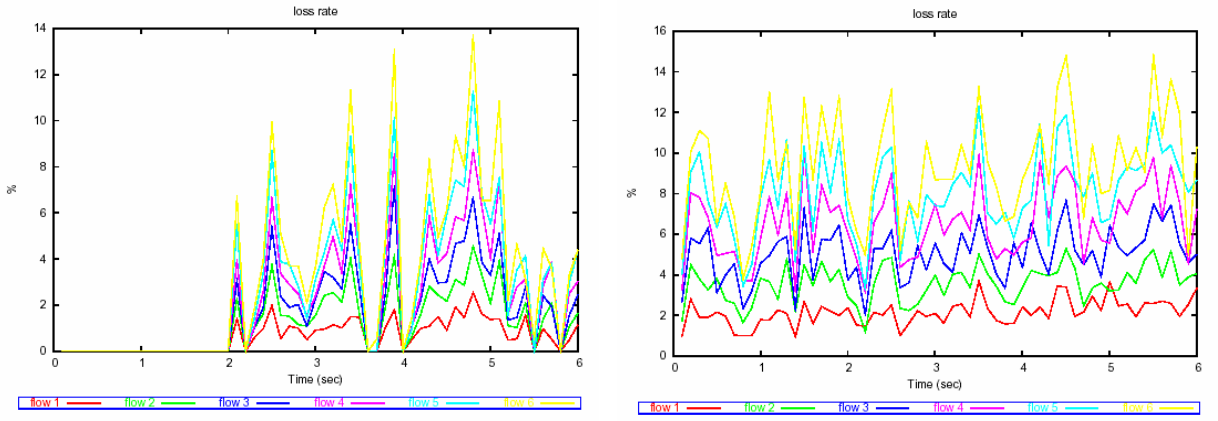

Figure 39: Loss ratio with EDS (left hand) and EDSRED (right hand) with no router congestion.

Figure 39 shows the differentiation performed by EDS and EDSRED when the router was not congested and it can be seen that EDSRED achieved a better differentiation than EDS. With EDSRED the differentiation occurred all times whereas with EDS the differentiation was more variable to the extent that sometimes there was no differentiation at all.

In summary, the RED version of the EDS implementation resulted in a smoothed loss rate and average length of the queue evolution and demonstrated that EDSRED achieved differentiation even when the router was not congested.

4.3 EDS TRANSPORT LAYER DESIGN

On a simple best-effort network, there is no way to control either the end-to-end delay or the loss rate. Packets are forwarded with both the delay and loss probability dependent upon the network load. EDS has been designed to reflect the IP design philosophy with respect to the plane of performance

differentiation. The same way that TCP has been designed to provide reliability on top of the unreliable IP network layer, three transport protocol have been designed which provides specific soft quality of service properties to applications to operate over EDS. These are defined as

- The RT-TP over EDS protocol ensures "as best as possible" end-to-end delay and a relative reliability to a real-time application.
- The SM-TP over EDS protocol ensures "as best as possible" end-to-end delay bound to reliable short message transport.
- The LM-TP over EDS protocol ensures "as best as possible" an improved end-to-end delay to bulk data transport.

NS simulation and real tests in emulated test-beds have demonstrated that this architecture is able to improve the flow specific performance criteria in the context of a realistic mix of heterogeneous traffic.

4.3.1 End-to-end delay constrained transport protocol over EDS:RT-TP

Assuming that the traditional best-effort service is replaced with the EDS service classes, a transport protocol then has the opportunity to use a specific best-effort class with a delay which is proportionally higher or lower than that obtained from the traditional best-effort service. This is known as the RT-TP protocol. The end-to-end delay is still dependent on routers load, however, by switching from one class to an other, the transport protocol can have some greater control over the end-to-end delay.

Consider an application that sends packets at a given rate with the expectation that the packets are received within a delay shorter than a known delay bound, and moreover, the end-to-end reliability has to be relatively high in comparison with traditional services. The RT-TP protocol includes as parameters both a delay bound value and a maximum loss rate to meet these requirements.

4.3.2 Interactive reliable transport protocol over EDS: SM-TP

The AIMD algorithm has been adapted for reliable and interactive short message transfer over EDS using SM-TP. This protocol is aimed at the application that has no hard delay requirements and can handle delayed file transfers, however with the need to maximize the probability that the transfer is complete in a short time.

The message length bound is assumed to be four packets and using the traditional TCP slow-start, the transfer of four IP packets takes at least three round-trip times (one packet is sent, then two, then one). If a packet is lost, the packet is re-transmitted and the transfer takes some additional round-trip times to complete. In order to minimize the completion time, the SM-TP starts with a window larger than four packets into which the entire message can fit. It is thus possible that the message is received in one round-trip time.

SM-TP uses the class with the highest drop probability while sending the initial burst. If the network load were high, packets would be lost but because of the use of loss rate differentiation a smaller burst is used. If the network load were low, a four packets burst would not increase the overall network load excessively particularly where the network has been sufficiently well provisioned. Moreover, the connection would not continue in slow-start as all packets have been sent.

Re-transmitted packets use the slowest class, where drop probability is low. Thus, in the case of low network load, the transfer is likely to be complete in a very short time while in the case of high network load packet loss should be expected.

SM-TP uses a class where the loss rate is lower taking the chance to complete a transfer in an acceptable time. The optimized protocol allows the user a relatively high probability of completion in a short time as a result of the burst at the beginning of the connection. However, the protocol uses the

class with the highest loss probability in order to lose packets where the network load is high, then, assuming some packets were lost, it continues using a slow class with a lower drop probability in order to increase the probability of transfer completion while the network load condition remain high.

4.3.3 Bulk reliable transport protocol over EDS: LM-TP

This strategy developed for the LM-TP protocol leads to a diminution of both the number of timeouts and the standard deviation of the performance of multiple connections.

TCP uses a well known algorithm that regulates its congestion window which provides some interesting properties in terms of efficiency and fairness between multiple connections. The greater the number of RTT exchanges a connection is able to complete without detecting a loss, the more the connection is effective with respect to data transfer. This is achieved as it manages to increase its window size to a higher value than a connection that has experienced loss. When a connection experiences a loss it needs to be temporary protected because of the case where either the re-transmitted packet is also lost or a second primary loss is experienced, throughput is severely reduced.

The marking strategy of LM-TP that consists of moving from one class to another may lead to packet re-ordering. Packet re-ordering can result in erroneous loss detection and thus to unnecessary packet re-transmission. The fast re-transmit optimisation has been modified so that it does not operate when selective acknowledgements (usually indicating a loss) acknowledge quicker packets.

4.4 LM-TP OVER EDS IMPLEMENTATION AND EVALUATION

4.4.1 Tests Methodology

The test bed consisted of two senders A and B, one router R and one receiver C linked by a path having as a bottleneck a 100Mbps Full Duplex Ethernet link.

The following tools have been used during the testing,

- **Mgen** to generate UDP flow in order to overload the router.
- **Iperf** to generate SCTP flow. The version used was patched to manage SCTP. [The patch was produced by Asim Iqbal (CERN) and subsequently improved by Marc Herbert (SUN/INRIA)]
- **tc** to configure of the router. [An add-on was provided to manage EDS by Pierre Billiau (INRIA)]
- **nistnet** to add delay between the senders and the receiver.

4.5 LM-TP EVALUATION

All adaptive packet marking algorithms have been implemented in the SCTP NS module with the LM-TP protocol implemented as an SCTP module in Linux. SCTP was chosen for its better implementation both in LINUX and NS in comparison to the equivalent one of TCP. The implementation as a module in Linux facilitated the test.

In Figures 40 and 41, the following conventions are used,

- IP: the default configuration of the router;
- red: a RED router : red limit 100000 min 6000 max 80000 probability 0.1 bandwidth 100000 avpkt 1065 burst 30;
- no diff: a EDSRED router performing no differentiation : Same as RED more hist_depth 1200 nb 6 fact_delay 1 1 1 1 1 1 fact_loss 1 1 1 1 1 1; and
- diff: a EDSRED router performing equivalent differentiation in delay and loss rate : Same as above but fact_delay 10 12 15 20 30 60 fact_loss 60 30 20 15 12 10.

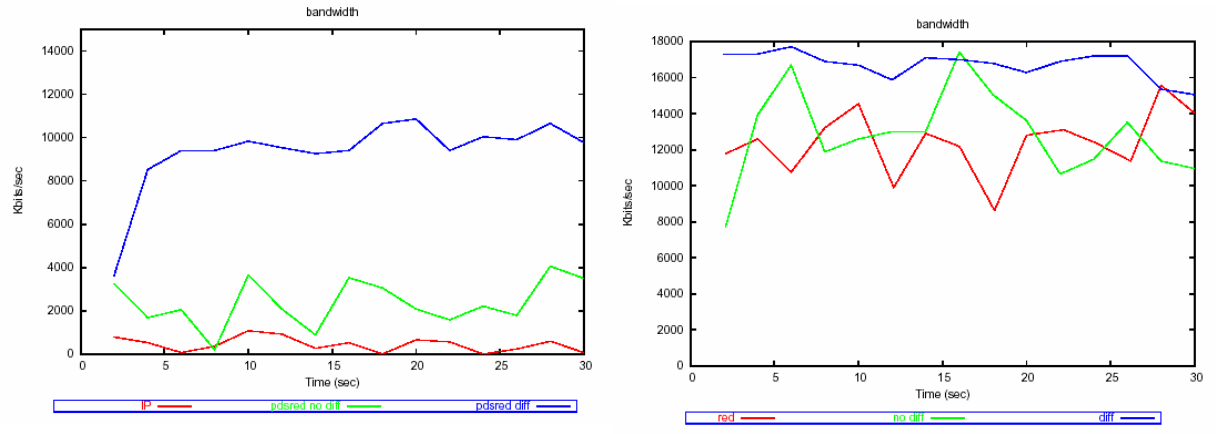

Figure 40 : Bandwidth of LM-TP (left hand) and TCP (right hand) over different routers when there is no delay.

When there is no delay and no overload, LM-TP achieved the same performance as TCP at 94Mbps. However when the router is congested (Figure 40) TCP achieved slightly better performance while LM-TP achieved a much greater bandwidth when routed with EDSRED performing the differentiation. This demonstrated that LM-TP is able to take greater advantage of the differentiation than achieved by TCP.

Extensive tests with the Nistnet emulator have shown that LM-TP over EDS is resistant to unfriendly UDP flows over short or long paths. Throughput obtain with lkSCTP, SCTP-lm and TCP have been

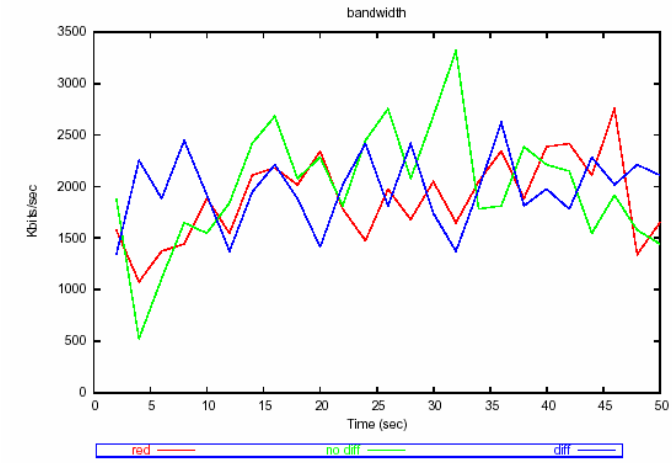

Figure 41 : Bandwidth of LM-TP over different routers when there is a delay of 200ms

compared in different conditions (load, delay) on different router configuration (IP default, RED, EDS, EDSRED). Where there is no delay the performance of SCTP-lm on EDSRED is comparable to that of TCP and moreover all the protocol showed their best performance on EDSRED.

However, when the delay was set to a high value (Figure 41) using nistnet (RTT=200ms), the protocols showed the same performance on all router.

4.6 CONCLUSIONS

In the testing reported here, it has been shown that SCTP-lm is able to take advantage of EDS where there is no network delay, but not where the network delay is high. However this is not of such significance because the aim and purpose of EDS is to improve the overall, global performance and not to improve performance of each co-operative flow separately.

Test plans are now being developed that will investigate how a mixture of heterogeneous flow, more closely corresponding to real traffic observed in Internet, may benefit from such an approach.

The SM-TP protocol will be also implemented during the next period and tested together with LM-TP over EDS.

In all these tests, the EDS level mechanisms have been implemented in software routers as no commercial routers currently support these new mechanisms. However, such software routers are very easy to deploy at the edge of the WAN where bottlenecks are often localized. Future work will be performed to evaluate the EDS mechanism particularly at the higher data rates available in the DataTAG provision.