Project Number: IST-2000-26417
Project Title: GN1 (GÉANT)

Deliverable D9.9

Experiments with Less than Best Effort (LBE) Quality of Service

Deliverable Type: PU-Public
Contractual Date: 31 August 2002
Actual Date: 30 August 2002
Work Package: WI 8.7
Nature of Deliverable: RE - Report

Authors:

Tiziana Ferrari (INFN/DATAGRID), Tim Chown (University of Southampton and UKERNA), Nicolas Simar (DANTE), Roberto Sabatino (DANTE), Stig Venaas (UNINETT), Simon Leinen (SWITCH)

Abstract:

In this document we present the work done to date and planned future work of the GEANT TF-NGN working group studying the potential benefits (to end users and NREN operators) and feasibility of deploying a Less than Best Effort (LBE) service on GÉANT and NREN networks. We present a brief overview of the evolution of the case for LBE, through the IETF diffserv WG and the Internet2 QBone project, and then describe some proposed scenarios for LBE deployment. Two sets of trials have been identified, one for implementation of LBE on the GEANT (and NREN) backbones, and one where LBE is applied at the edge networks where congestion is more likely (the latter case requires DSCP transparency end-to-end). Some encouraging results from the former case are presented, and proposed trials for the second case are also described.

Keywords:

IP Quality of Service, Less than Best Effort (LBE), Scavenger

Table of Contents

1 INTRODUCTION 4

2 LBE SERVICE DESCRIPTION 6

3 APPLICATION SCENARIOS 8

3.1 MIRRORING 8
3.2 PRODUCTION AND TEST TRAFFIC 9
3.3 SUPPORT OF NEW TRANSPORT PROTOCOLS 10
3.4 TRAFFIC MANAGEMENT FROM/TO STUDENT DORMITORY NETWORKS 10
3.5 NETWORK BACKUPS 11
3.6 ESTIMATION OF AVAILABLE BANDWIDTH 11

4 LBE TEST PROGRAMME 12

5 PLANNED TRIALS FOR SITE LBE POLICY IMPLEMENTATION (ACTIVITY 2) 13

6 EXPERIMENTAL RESULTS OF TESTS PERFORMED ON GÉANT 15

6.1 TEST EQUIPMENT AND NETWORK INFRASTRUCTURE 15
6.2 LBE AND BE PERFORMANCE MEASUREMENT WITHOUT CONGESTION 17
6.2.1 Packet loss and throughput 17
6.2.2 One-way delay 17
6.2.3 Instantaneous packet delay variation 18
6.2.4 Out-of-sequence packets 19
6.3 LBE, BE AND IP PREMIUM PERFORMANCE IN CASE OF CONGESTION 20
6.3.1 Packet-loss 20
6.3.2 Throughput 20
6.3.3 One-way delay 21
6.3.4 Instantaneous packet delay variation 23
6.3.5 Out-of-sequence packets 23
6.3.6 Conclusions 25

7 GÉANT ROUTER CONFIGURATION 26

7.1 THE M-SERIES QUEUING ARCHITECTURE 26
7.2 GÉANT CONFIGURATION BEFORE THE TESTS 27
7.3 TEST CONFIGURATIONS 27
7.3.1 First configuration 28
7.3.2 Final configuration 29

8 CONCLUSIONS AND FUTURE WORK 30

9 ACKNOWLEDGMENTS 31

10 REFERENCES 32

11 ACRONYMS 34

A1. ANNEX 1 – ROUTER CONFIGURATION 35

A.1 LAST (THIRD) GÉANT ROUTER CONFIGURATION 35
A.2 FIRST GÉANT ROUTER CONFIGURATION 36
A.3 SECOND GÉANT ROUTER CONFIGURATION 37

Executive Summary

In recent years there has been a growing demand by users in the European research community for a high quality of service for the applications they are running. The most common approaches to delivering the performance that the users require are either to increase the network provision in advance of demand (a technique commonly referred to as "overprovisioning"), or to deploy some kind of "Better than Best Effort" quality of service mechanism in the network (e.g. the Premium IP service as defined in GÉANT Deliverable D9.1).

However, overprovisioning can only be applied where the funds of the network operator permit. Thus while GÉANT enjoys 10Gbit/s connectivity between many of its PoPs, and a number of NRENs have similar capacity backbone networks, such provision is far from universal. For example, while the UK core network (SuperJANET) operates at 10Gbit/s, many universities are "only" connected at capacities in the order of 155Mbit/s, and many further education colleges have 2Mbit/s links. While it is expected that Premium IP deployment can offer a good service for many users (whose own view of "quality of service" may vary), widespread deployment can be a far from trivial exercise where multiple administrative domains are involved, dynamic bandwidth brokering is required, and aggregate reservations have to be considered.

In this report we evaluate a different approach to quality of service, where a Less than Best Effort (LBE) service is defined. Intuitively, one might think that very few users would be willing to run applications that would receive a worse service than regular Best Effort (BE) traffic. However, we argue that in providing a traffic class that can expand to utilise the available bandwidth on a link, without any significant effect on the BE (or Premium IP) traffic, a number of new network usage scenarios can be met. The general principle of LBE is that in the presence of congestion, LBE packets are always dropped before BE (or better) packets.

In this report we begin by giving a service description for LBE. In this description we choose to be interoperable with the Internet2 Qbone Scavenger Service (QBSS) by using the same diffserv code point (DCSP) value of 001000. By using a common value we enable LBE between European and Internet2 research sites, and we can extend interoperability further if other networks (e.g. in Japan) use that same DSCP. We also seek to use a similar minimum bandwidth guarantee for LBE, to help prevent network starvation for TCP LBE applications. We then list example scenarios for LBE deployment including data mirroring, GRID data transfers, non-invasive test traffic, network backups, student dormitory networks and (potentially) measurement of available network bandwidth on a network path. LBE offers a kind of overlay network that generally allows high volume, low priority applications to run in the available bandwidth without adversely affecting regular BE traffic.

We demonstrate an LBE service implemented on the GÉANT backbone network on which we vary parameters including the weights and priorities for LBE, BE and Premium IP queues. The results show that with appropriate priority selection an LBE service can run without any notable disruption to the BE service, and with minimal packet reordering in the presence of congestion. We also describe future tests that will be run to evaluate LBE where used in end site (university) networks across NREN and GÉANT backbones, where LBE drop policies are only applied at the network edge and the intermediate routers merely offer DSCP transparency (they preserve the LBE DSCP).

The results obtained to date suggest that an LBE service can be deployed on GÉANT and NREN networks, and that both the NRENs and their end users can benefit. One of the attractive features of LBE is that it can be deployed (in terms of a queuing and drop policy) incrementally on a network path, with other routers only needing to offer DSCP transparency. It must be noted that LBE is not a panacea for quality of service. For example, in the presence of a link that is continuously congested it offers no benefit if none of those applications using the bandwidth can be run as LBE applications. Nor does it give the high quality delivered by Premium IP. However, we feel it is a service that may offer notable benefits in a number of application scenarios.

1 INTRODUCTION

Between 1999 and 2002 there have been several initiatives to design and implement services capable of offering guaranteed and predictable network QoS to end-users. The Internet2 Qbone initiative [qbone] started in 1999, whilst within GÉANT work for the definition of a Premium IP service started a little later, in November 2000 [prem-ip]. The Internet2 Qbone and GÉANT initiatives experienced rather different levels of success: the GÉANT work lead to the definition of Premium IP service, based on the diffserv Expedited Per Hop Behaviour (EF-PHB) capable of operating in a multi-domain environment [geant-d91]. The Internet 2 QoS Working Group [i2-qoswg] has put development of Premium IP [i2-prem] on hold, and is instead looking at other alternatives to promote quality in application experiences for its users. It is worth noting that:

- The Internet2 Qbone initiative started earlier than the GÉANT initiative. Of the problems encountered they cite the need for all-or-nothing network upgrades for providers, dramatic changes to network operations, peering arrangements, and business models, and the absence of suitable means to verify service. Lack of router functionality on the Cisco routers was also an issue at the time;

- The networking environment in which the Qbone was operating, Abilene [abilene], is totally different from that of GÉANT. Although GÉANT is currently the highest capacity research network in the world, there is still a degree of diversity of capacities available in the network. There are still locations connected at capacities of less than 622Mbps. In the case of Abilene, the network is much more homogeneous, with Gbps capacities available almost everywhere. This shifted the focus of QoS onto issues such as quality of fibre connections, stability of routing, quality of the end systems and so forth. The elimination of such problems now forms the focus of the Internet2 End-to-End Performance Initiative [e2epi]). Also, Internet2 is evaluating the Alternative Best Effort (ABE) service [abe].

The original ideas for an LBE-like service arose from work in the IETF diffserv WG [diffserv]. However, it is the work of the Scavenger team that has helped raise the profile of such a service, to the extent that adoption is now being evaluated for GÉANT and by NRENs. There is now a new Internet Draft by Bless et al [bless-le] defining a similar service. Naturally this initiative will be tracked and fed into where appropriate by the TF-NGN LBE group.

The main idea of LBE is that this traffic class is able to make use of unutilised bandwidth in the network, but in a way such that in cases of competition for resources, the LBE traffic will be discarded before any Best Effort (BE) or higher-class traffic. Therefore, the LBE traffic class is subject to relatively high risk of high packet loss. In terms of implementation, it is based on a sub-set of the techniques used to deliver guaranteed and predictable network QoS, more specifically the scheduling mechanisms. The LBE service has been met with enthusiasm from users, and several useful application scenarios exist for this traffic class. These are described in more detail in Section 3.

To summarise some of the application scenarios, LBE offers the ability to users that have demands for high volume transfers but no strict time constraints to transfer their data to make use of all unutilised network resources without interfering with higher priority (including regular BE) traffic. Traditionally, many network-conscious users would refrain from using the network during normal working hours because of the fear of interfering with the users at large, and would therefore schedule their activities to run at night. LBE allows these users to use the network at any time of the day. Specific scenarios include FTP mirroring, GRID data transfer, experimental data transfers, network backups, control of student dormitory network traffic, and a possible way to estimate available bandwidth on a link in a non-disruptive way.

The LBE traffic class can use bandwidth that is not used by higher priority traffic using the standard Best Effort traffic class, and in cases of competition for resources, these will be given to the normal users and the LBE traffic class will suffer packet loss. Obviously, users of LBE must be tolerant of packet loss in order to make use of the service (the service description suggested in Section 2 of this document, and as also defined in QBSS, allocates a small percentage of bandwidth as a minimum guarantee for LBE to avoid complete starvation for the service).

There are of course other potential incentives to use LBE services, other than the will of network conscious users to be "friendly" to the network. One of these incentives is related to billing in that the LBE service could be charged at lower rates than the normal Best Effort service, which will encourage its use. Having more users or applications using LBE will ensure better QoS to the standard Best Effort service, and therefore the community at large can enjoy better network performance. On the other hand, network providers have an interest in charging less for LBE because if users make use of it, there is less need for network upgrades, which implies less expenditure in terms of hardware and connectivity.

On GÉANT the main focus remains that of ensuring predictable and guaranteed QoS, mainly because of the heterogeneity of the network connectivity on an end-to-end basis within Europe. However, the LBE service does have useful application scenarios and therefore will be supported by GÉANT and some NRENs. Note that the LBE service can be implemented incrementally and it does not require peering networks to support it, as no end-to-end guarantees are offered by LBE. In particular, in June 2002 during a meeting between the GÉANT community and Internet2 representatives it was agreed that:

- The LBE traffic class should be identified by a globally unique DSCP value (DSCP=8 in order to comply with the DSCP adopted by Scavenger), and treated at least transparently within GÉANT. This means that the DSCP value of 8 should be preserved within GÉANT although no differential treatment may be applied.

- The NRENs interested in LBE would test and deploy the service, in collaboration with Internet2, utilising either circuits between the NREN and Abilene or the circuits between GÉANT and Abilene, depending on how this connectivity is implemented for each NREN.

In fact, GÉANT will go a step further than being transparent to LBE, and will provide differential treatment according to the specification of LBE provided in this document.

The rest of this deliverable will outline application scenarios of LBE and a description of the tests carried out on GÉANT. To date, no tests have been performed between NRENs and Internet2 for LBE, but such tests should occur in the near future. It will also be desirable to identify LBE usage scenarios to networks in Asia, including in Japan and Korea. It would naturally be good to encourage use of the same globally unique DSCP value there, just as between Europe and Internet2.

Online information can be found on the TF-NGN LBE WG web site [tfngn-lbe].

2 LBE SERVICE DESCRIPTION

The definition of the LBE service follows the basis that a given differentiated services code-point (DSCP) is used to convey the meaning that packets bearing such a DSCP value can be given a lesser service than regular best effort (BE) traffic. If congestion at a given interface is produced by LBE traffic, then congestion is completely transparent to packets belonging to higher-priority classes like IP Premium and BE. Congestion is produced by LBE traffic if the output capacity of the interface is exceeded because of the injection of LBE packets, but both the instantaneous and average BE traffic rate can be handled by the interface without introducing BE packet loss. Performance of packets belonging to higher-priority classes cannot be protected against congestion if the amount of traffic belonging to that class or to higher-priority classes at a given output interface is sufficient to produce either short or long-term congestion, regardless of the presence of LBE packets. Protection of higher-priority classes from LBE traffic has to be supported both with and without congestion. Protection requires that packet loss, one-way delay, Instantaneous packet delay variation and throughput of streams belonging to higher-priority traffic classes should not be affected by the presence of LBE traffic, either with or without LBE congestion.

No end-to-end guarantees are provided to flows adopting the LBE service. This means that the LBE service is not parameterised, i.e. no performance metrics are needed to quantitatively describe the service. In addition, no seamless end-to-end service is provided by LBE. This implies that LBE can be supported incrementally on congested interfaces as needed without requiring any LBE service support in peering networks.

In order to avoid starvation of TCP-based flows based on the LBE in case of congestion, a very small portion of the available bandwidth should be offered to LBE as a minimum throughput at routers implementing the LBE policy (on Scavenger, this value is set at 1%).

For routers not implementing the LBE queuing and drop policy, the minimum requirement is that the DSCP value chosen to indicate the LBE service is passed transparently across the network (not set to another value or cleared to zero). In routers not implementing LBE drop policies, it is expected that LBE receives the same queuing priority as BE. Note that in routers that have IP Premium implemented, we would expect LBE to be also implemented. Since the support of QoS is particularly important at congestion points, we expect different traffic classes, like IP Premium, BE and LBE, to co-exist at network bottlenecks. It is however essential to verify that the proper co-existence of the three different services can be achieved.

DSCP transparency is far easier to implement as a diffServ service than end-to-end Premium IP. Aggregate bandwidth reservations do not need to be made. Rather, a network can enable LBE by default, if it does not alter the DSCP value, then deploy incrementally the queuing mechanism where needed, for example in points where congestion is more likely to occur (current this is most likely at the edges of networks).

The agreed DSCP for LBE is the same DSCP as used for Scavenger on Internet2, namely binary 001000. It is recommended that all routers on the GÉANT network and the attached NRENs at the very least do not reset this DSCP value, which then allows edge sites and MANs (where congestion is currently more likely to occur) to incrementally implement the LBE drop policy. While it is possible that an LBE-tagged packet may traverse all but the last hop into its target network before being dropped, that apparent "waste" of bandwidth is compensated by the resultant TCP backoff in the LBE application resulting from the packet loss. In any case, such “waste” will affect only routers and links whose resources would have been left unused by users of other traffic classes.

Protection of non-LBE classes from LBE congestion can be achieved by placing LBE traffic in a dedicated queue for any output interface that is subject to congestion. It is suggested that the LBE queue is not shared with other traffic classes, since the presence of large LBE bursts could have an impact of the queuing delay experienced by packets belonging to other classes waiting for the long LBE bursts at the head of the queue to be transmitted.

Different scheduling algorithms can be adopted to service the LBE queue and the higher priority queues enabled on a given output interface. In case of algorithms requiring a bandwidth share assignment to each configured queue – such as Weighted Round Robin and Weighted Fair Queuing - it is recommended that a very small bandwidth share be assigned to the LBE queue. The most appropriate configuration is an implementation issue that depends on the specific router platforms in use, on the number and type of QoS services enabled on a given interface and in general on the network set-up.

Note that the configuration of a LBE queue and its corresponding bandwidth assignment need not be made on routers that are purely DSCP-transparent. It is only made on routers implementing LBE drop policy to ensure TCP applications adopting the LBE service can back off in some sort of predictable fashion, rather than be starved of all bandwidth and above all, to protect higher priority classes from LBE congestion. In essence this creates an "LBE overlay network" that can grow to occupy unused bandwidth.

Usage of LBE may be done voluntarily, or by site policy. It is expected that LBE marking will be performed in the end host system voluntarily, or at site border routers by enforced policy. The former case may by example be an FTP server that recognises an extension whereby the user can request LBE to be applied to the file retrieval (see the FTP mirroring scenario in Section 3). The latter case may for example be applied to a student dormitory network. Note it is the sender of the data that must mark the traffic, thus in the case of a user wanting to download a file using LBE, who wants to be "network friendly", it is the server that has to mark the traffic returned to the user. This implies the requirement for a signalling system or of some statically defined mutual agreement on the traffic class needed by user groups.

The Less than Best Effort (LBE) service is defined in our work in a very similar way as the Internet2 QBone Scavenger Service (QBSS). In fact, while the precise implementation method may vary, we intend to make the Scavenger-LBE implementations interoperable, in the sense that the service can be used meaningfully between European and Internet2 end sites and users.

3 APPLICATION SCENARIOS

In this section provide an overview of a number of scenarios where the use of LBE may be beneficial to a user, a network provider or both. This is not intended to be an exhaustive list of scenarios, but gives a flavour of what can be achieved if providers support the LBE service for their communities.

3.1 Mirroring

Content on the Internet that is accessed by a large number of hosts, which in turn are located at many different places, is often replicated in several locations on the Internet. Users can then retrieve content from a replica or mirror site that is topologically close to them, which ensures more efficient access and imposes less resource utilisation on the server with the master copy of the content (to the extent that in some cases the master will only be accessed by replicas). In turn, this enables the provider of the content to serve a larger community of users.

One way to manage the replication of data in the network is to simply have a number of caches spread around the network, where caches have their own copies of data that has recently been accessed or are in high demand. A standard cache will only request data from the master when a client accesses it and it will typically need the data as quickly as possible. This can be thought of as synchronous access. Another commonly used technique, especially for data that changes infrequently, is that of asynchronous updates, where changes at the master are propagated to the replicas at regular intervals independently of when clients access the data. This is also called mirroring and in this case timing is usually not critical. Ideally the update should not penalise users accessing the data, in the sense that users should observe the same performance in terms of transfer rates, latency etc. during the update as they normally would experience. In order to achieve this, the update will either be done at fixed times that are expected to be off-peak periods, or master and replica might perform updates only when the load is below a certain threshold. By using LBE, updates can be done at any time of the day without penalising the user data traffic. This is very useful for large updates that might consume a lot of bandwidth and last for a long time. Not only can one avoid penalising user data traffic for that particular service, but also other data traffic in the parts of the Internet where these packets are treated as LBE. This might be on the local network where the master or replica is located, but can also be larger networks, one site or an operator’s network.

Mirroring is commonly used for popular FTP sites. Some examples are mirroring of distributions of Linux, NetBSD etc (from the various distribution makers) and RFCs from ftp.ietf.org. Such mirroring is often done by a nightly job at the mirror site that contacts the master FTP server, compares the remote and local files, retrieves any files that have been modified, and deletes files that are no longer present. Traditionally such mirroring is done using FTP, but other mechanisms like rsync and cvs using SSH for authentication are also used to some extent. By using LBE for the updates, one can make sure user access to the FTP servers is not penalised. For FTP sites such as the above, there is another good reason for using LBE. Mirror site providers rarely make any profit by offering such services, and with the use of LBE for mirroring (and potentially user downloads as well) one can make sure the more business critical use of the network is not harmed.

In this application scenario, it is the sender of the traffic (the master) that needs to classify as LBE all packets for the receiver (the replicas), but the request for the traffic comes from the replica. This leads to an important implementation issue. If the master knows the replicas, the master can distinguish between replica and user access, and classify only packets destined for replicas as LBE. This requires some detailed configuration management, and there are also cases where the master may not be able to distinguish between user and replica access. Ordinary users also have different requirements. An ordinary user may not need the data immediately, and prefer LBE to favour other data traffic, or maybe to save money if the Internet provider charges LBE differently.

The user may not be aware of LBE whilst the user's applications or the operating system may somehow decide whether LBE should be used. This implies that in the case of downloads, it is the receiver that knows best whether LBE should be used, whilst it is the sender that has to mark the packets (receiver denotes the party receiving the downloaded data). Because of this, we need a mechanism that allows the receiver to signal the sender that it should use LBE to send the traffic to its destination. One possibility is that the replica (or user) uses LBE to send the request to the master, in which case the sender uses LBE to send traffic if data packets from a replica or user are marked as LBE. Another approach is to signal LBE at the application level. For instance with FTP, it is possible to add a command (SITE) to the FTP protocol that enables the receiver of traffic (user or replica) to signal to the master to send the traffic using LBE.

3.2 Production and test traffic

The LBE service can be effectively used for protection of high-priority traffic from low-priority traffic. For example research centres involved in experimental exchanges of large data volumes or in testing of new applications/middleware, like the GRID community, may be interested in protecting high-priority production traffic from potential congestion produced by test packets. High volumes of data may be replicated in multi-tier caches, an example being experimental data distributed to tier sites from CERN. Where real-time delivery is not critical (i.e. "just in time" delivery is acceptable), LBE can be considered for the delivery. The service is particularly interesting when the link providing access to the NRN infrastructure or one or more interfaces within the local area network are subject to congestion.

Data management is one of the fundamental functionalities of most experimental computational grids. The development of data management middleware requires testing of a large set of functionalities like data replication, code and data transfer (job submissions), data communication for distributed applications, databases operations, directory related messages, etc.

Testing of database replication can greatly benefit from the use of LBE when copying very large amounts of data from a given source site (e.g. from a Tier 1 site) to multiple remote destination sites (e.g. Tier 2 and Tier 3 national sites) [RC]. In fact, data replication sessions need to be sufficiently frequent to grant Computing Elements an efficient local access to large portions of data in order to minimise data access latencies and to avoid communication bottlenecks at given Grid sites.

LBE can be effectively used both in the "push" and in the "pull" approaches to replication. The push approach requires write access to remote sites. Conversely, in the case of pull, it is the remote site that requests a given file to the upper-level file server; this "on-demand" approach does not imply synchronisation between remote sites. The interested reader can find more information about results of LBE testing for the support of GRID middleware and applications in [EDG7.3].

In many cases test traffic can be easily classified and marked with the LBE DSCP, for example on the basis of the source and destination IP address, especially when test equipment resides in dedicated subnets. Then, an appropriate scheduling algorithm protecting BE traffic from the previously identified test packets has to be enabled in each potential congestion point. The support of the LBE service can be adopted in a customer network independently from the availability of LBE support in the wide area networks connecting remote test sites. In this case, the LBE DSCP set by the customer network has to be preserved during packet forwarding.

LBE can be similarly used to protect production traffic from TCP sessions used for the exchange of very large files (for example in database replication). However, in this case it has to be reminded that the long TCP sessions based on the LBE service are possibly subject to higher packet-loss rates than plain BE traffic. Packet loss can have a negative impact on TCP performance and link capacity utilization, especially when high-speed long-delay links like the transatlantic connections are involved in the data exchange.

3.3 Support of new transport protocols

One of the main concerns of high-performance applications and middleware is the efficient utilisation of network capacity when long-distance high-speed links are used. For example, many of the middleware components of distributed systems, like authentication, database replication and the exchange of jobs and input/output data, require reliable high-speed communication among remote grid nodes (e.g., computing elements, storage elements, resource brokers and information servers). The efficiency of communication on WANs is fundamental to guarantee the reliability and robustness of GRID computing, especially of the middleware components based on the exchange of extremely large amounts of data. High-performance transmission over long-distance connections is essential for the support of GRID-based applications in a large number of scientific areas like high energy physics, bio-informatics and earth observation. The capacity of high-speed links can be efficiently used by long TCP sessions only if TCP socket sizes are properly tuned according to the bandwidth-delay product of a given flow and if the stream does not suffer from loss. In case of packet drop, the traditional congestion control and avoidance algorithms in TCP can severely limit the TCP performance given the long time needed to recover after the loss event to bring the congestion window size back to its original optimal value. For this reason, several TCP extensions that improve the protocol efficiency and also alternative new transport protocols are under study and definition by the research community. Normally, such alternative protocols react to congestion more aggressively than TCP.

The co-existence of applications based on traditional TCP stack implementations and applications that will adopt such new transmission algorithms have to be guaranteed. The LBE service could be used during the test phase of new transport protocols, or even in production, to protect TCP-compliant traditional traffic from the test applications using the more aggressive non-TCP compliant transmission techniques.

3.4 Traffic management from/to student dormitory networks

One of the deployment scenarios for the QBSS on Internet 2 has been in student dormitory networks at university sites. The premise is that student traffic from living quarters that is passing to the Internet is generally deemed lower priority than staff and research traffic (e.g. students running peer-to-peer transfers should not detrimentally affect response from the university's web server for external visitors). While students should be able to reach campus facilities, facilities off-campus may be deemed more “expendable” when congestion is occurring at the site's Internet access router.

In such cases LBE marking can be applied on routers connecting the dormitory networks to the campus network. Some or all traffic can be so marked. Note though that where students are downloading data to their rooms the received data is not LBE-tagged unless the server side application specifically honours the DSCP seen on incoming requests. As this is unlikely, the major impact of LBE-tagging would be on data being exported from room networks, e.g. peer-to-peer file transfers, or FTP servers run in student networks.

In this scenario the LBE-tagged traffic (at the campus-dormitory border) may never leave the university network if dropped at the campus-Internet border. However, this is not always the case, and the LBE tagging may be useful on ingress to the target network (if that is the other point on the data path most likely to have congestion); thus we should argue to apply LBE rather than purely using site-specific marking and dropping.

Note that it is generally the responsibility of the university in question to ensure that the type of data that flows from student dormitory networks meets their NREN’s AUP when that traffic flows out through the NREN network (as it is for any traffic leaving the university via the NREN network). The university can apply its own security policies (including firewall filters, or perhaps use of IPv4 NAT) to determine what protocols are accepted to devices inside the student dormitory network (e.g. some peer to peer applications may be blocked). Currently peer to peer applications have something

of a bad reputation through abuse of copyright via services like Napster and Gnutella. However, given a student is very likely to have their own PC, quite probably always-on, in their study room, it is reasonable to expect that they will wish to access data on that machine (e.g. via ssh, (s)ftp or http), or run more peer to peer applications to that machine, whether present at it or not. Allowing such access, but at a lower priority through implementing LBE, may be a good compromise for universities where external bandwidth is limited.

3.5 Network backups

Data mirroring and GRID transfers are forms of data replication on a network, as are network backups, whether directed to remote tape or disk. The write-speed of new LTO drives can certainly saturate 100Mbit/s network links (e.g. 30MB/sec is typical). Thus in some topologies it would be useful to use LBE to do non-disruptive backups to remote servers or devices. It may be that a central backup facility is offered by a university where remote sites may dump to the central store, possibly even between sites. Rather than just running the dump at night, it could be run all day as well using LBE, raising the effective dumping throughput and capacity.

Such dumps may be typically done over a TCP ssh tunnel for added data security. The LBE tagging could be done by the local access router or potentially by modifying the ssh code used for the backup scripts.

3.6 Estimation of available bandwidth

It may be possible to use LBE-tagged traffic to gain some non-disruptive (to BE traffic) estimate of available bandwidth on a given link or between end points.[1] However, this area of study requires further work to prove its potential, and to do so in a way that can be shown to be non-disruptive.

Available Bandwidth has been identified as an important network performance metric in GÉANT Deliverable D9.4 [geant-d94], but no tools were found that were able to measure it. The use of LBE could fill this gap.

[1] This suggestion was made by Sylvain Ravot of the Datagrid project in conjunction with tests between the California Institute of Technology and CERN
[Ref: http://www.internet2.edu/henp/Ravot.ppt].

4 LBE TEST PROGRAMME

Some preliminary experimental activities will be carried out to prove the compatibility of the new service with the other existing services and the functionality set already supported by GÉANT. In particular, the test programme is composed by the following activities.

- Activity 0: proposal and discussion of the LBE service specification;
- Activity 1: study of the feasibility of the LBE implementation both in the GÉANT infrastructure and in the backbones managed by NRNs interested in the LBE service;
- Activity 2: LBE and BE performance analysis in customer-based scenarios according to which the LBE service is only enabled locally in customer networks. In this phase the transparent forwarding of the LBE DSCP is the only feature requested to the NRNs and to GÉANT;
- Activity 3: tests of the LBE service implementations where the service is supported by both a restricted set of customer networks and a set of NRNs and/or by the GÉANT infrastructure;
- Activity 4: tests of performance and interoperability of the LBE service with the QBone Scavenger service.

The feasibility of the LBE service in the GÉANT network according to Activity 1 has already been studied by defining the implementation and configuration of the service and by conducting preliminary laboratory tests of the features we intend to support in the production GÉANT network. The experimental results on the LBE service performance presented in Section 6 of this document (Activity 3) were extremely useful to compare and validate different configuration solutions.

Work related to Activity 2 is ongoing. Experiments will be carried out according to the test plan presented in Section 5.

5 PLANNED TRIALS FOR SITE LBE POLICY IMPLEMENTATION (ACTIVITY 2)

In this section we describe proposed trials of the use of LBE as defined in Activity 2 of the test schedule.

In this case we are assuming the network congestion occurs at the end site access routers. As a consequence our scenario has the following features:

1. LBE policy is only implemented on the site access routers. All other routers on the path between the sites only provide DSCP transparency.

2. Bandwidth at the end sites is limited to 10Mbit/s (e.g. by use of switch devices to force FE router ports to negotiate down, or by manual configuration).

The proposed trials will be run between the University of Southampton – UoS – (UK) and UNINETT (Norway). The outline topology will be as illustrated below.

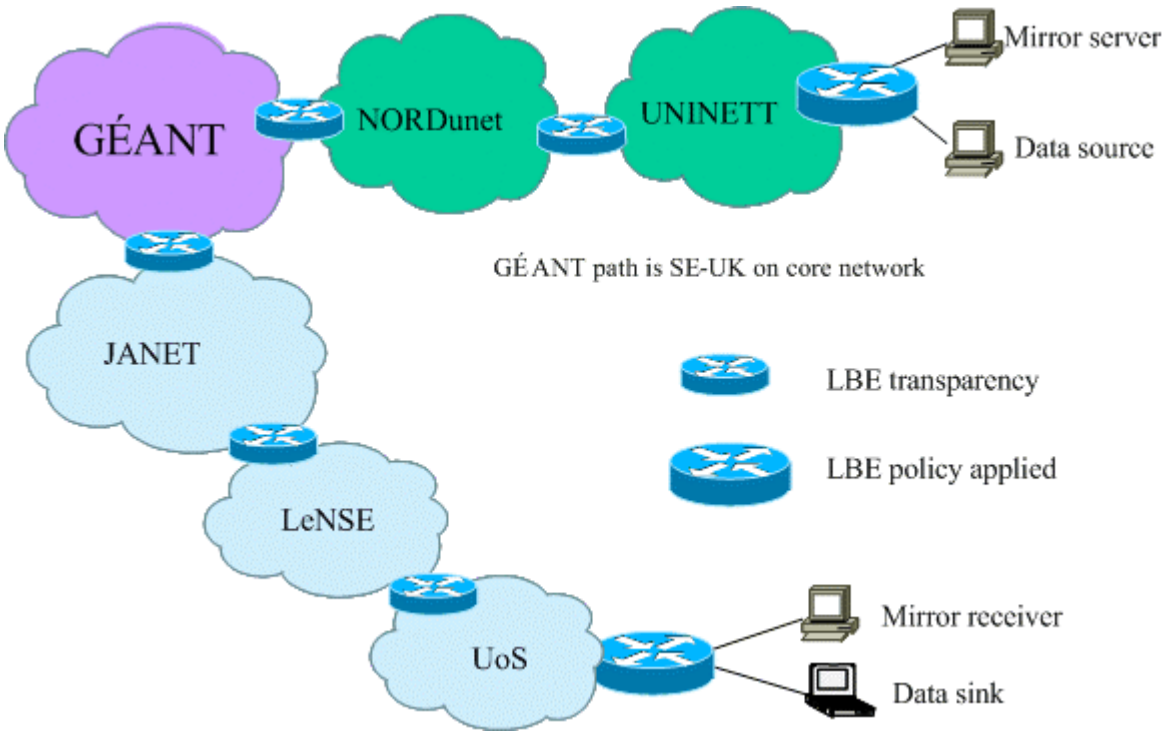

Figure 1: Outline network topology for end site LBE policy implementation trial with DCSP transparency on intervening MAN, NREN and GÉANT network.

Note that the UoS test site is connected via its own campus network to its regional MAN (LeNSE) and JANET to the GÉANT core network. The test nodes in Norway may be located at UNINETT offices, or at a university site (in which case the above topology will be extended).

Both UoS and UNINETT have used the NANOG traceroute utility to verify that DSCP transparency is in place between their end sites and the GÉANT core routers. In addition, DANTE has verified that the UK-SE (UK-NORDunet) link can be made DSCP transparent for the proposed tests, completing the end-to-end transparency.

The NANOG traceroute produces output like the following where the DSCP is altered (reset):

```
[nanog]; ./traceroute -t 32 www.dfn.de
traceroute to sadr.dfn.de (192.76.176.3), 30 hops max, 40 byte
packets
 1  swiCS3-V4.switch.ch (130.59.4.207)  1 ms  1 ms  1 ms
 2  swiEZ2-G3-3.switch.ch (130.59.36.18)  1 ms  1 ms  1 ms
 3  swiCE2-G2-2.switch.ch (130.59.36.41)  5 ms  5 ms  5 ms
 4  switch.ch1.ch.geant.net (62.40.103.17)  5 ms  5 ms  6 ms
 5  ch.it1.it.geant.net (62.40.96.34)  19 ms (TOS=0!)  19 ms  19 ms
 6  it.de2.de.geant.net (62.40.96.61)  28 ms  28 ms  28 ms
 7  de2-1.de1.de.geant.net (62.40.96.129)  28 ms  28 ms  28 ms
```

The UoS access router will be a Cisco 7200 series router. The UNINETT access router is likely to also be a Cisco product.

We propose to use the FTP mirroring scenario to evaluate the effectiveness of the LBE marking. The aim of the trials will be to fill the site access link with LBE traffic and to observe the effect on regular BE traffic between data source and sink hosts at the two sites, and to hosts on sites at external networks. We hope to demonstrate that the FTP mirroring can run continuously on available bandwidth without any significant effect on the regular BE traffic.

If SmartBits equipment is available, we will consider using it as part of the tests, and seek to use the same metrics for evaluation as are used on the trials of LBE implementation on the core network. Otherwise tools such as Chariot [chariot] will be used to assess the BE traffic properties.

There is interest in UKERNA (UK) in the results of LBE trials both within their QoS Development Programme and in testbed initiatives such as the Managed Bandwidth Next Generation project. We may also consider applying LBE in IPv6, where peer-to-peer applications may become more prevalent (with more devices, an extended address space, and no NATs impeding connections into networks). The 6NET project [6net] may be a good vehicle for such work.

6 EXPERIMENTAL RESULTS OF TESTS PERFORMED ON GÉANT

The support of the Less than Best Effort (LBE) quality of service was enabled on a subset of GÉANT routers in order to carry out preliminary test activities, whose goals are manifold:

- The understanding of the feasibility of the LBE service and in particular, the study of its compatibility with other services already supported by the infrastructure, namely, IP Premium and Best Effort (BE).
- The support of LBE DSCP *transparency* – as explained in Section 2 – for a restricted traffic class in a subset of the infrastructure, i.e. of the capability of forwarding LBE packets by preserving the integrity of the original DSCP carried by the packet at the ingress GÉANT router.
- The comparison in terms of effectiveness of different scheduling configuration solutions for the support of the LBE traffic class and, in particular, the analysis of the Weighted Round Robin algorithm available on the GÉANT routers M160 and of its effectiveness in providing isolation between LBE and the remaining higher-priority traffic classes: IP Premium and BE. The presence of LBE traffic in the network should be transparent to other classes both with and without congestion.

We have tested the co-existence of three traffic classes (LBE, BE and IP Premium) in different traffic load scenarios and have analysed their performance in terms of the following network metrics: packet-loss, throughput, one-way delay and instantaneous packet delay variation (IPDV). We have also investigated the extent of packet re-ordering and the effect this may have on end-to-end TCP performance.

Three different router configurations have been adopted during the test sessions for the tuning of two scheduling configuration parameters:

- the *queue weight* assigned to the BE and LBE queues to define the bandwidth share assigned to a given queue in case of congestion
- the *queue priority*, which can be set to *high* or *low*, as explained in Section 7 to determine the queue scheduling order during congestion, if both queues hold a negative credit.

Different queue weights have been assigned to the LBE queue. While both the (small) weights assigned correctly protect BE traffic from LBE congestion, the queue priority proved to be critical for the minimisation of packet reordering, as shown later in this Chapter. The router configurations adopted are available in Annex 1.

In the following section, the bandwidth percentages refer to the amount of traffic generated by the SmartBits STM-16 interfaces.

The total traffic sent is the sum of the traffic sent by the SmartBits generators with the production traffic exchanged on the test data path (currently between 50Mbps and 100Mbps) and additional test TCP traffic of approximately 210Mbps.

6.1 Test equipment and network infrastructure

A subset of the GÉANT infrastructure consisting of a set of STM-16 and STM-64 links and of the relevant terminating routers has been used to test the class of service mechanisms needed to support the LBE class. Different traffic scenarios have been generated by combining different transport protocols – UDP and TCP – and a variety of traffic loads and streams for each of the three above-

mentioned classes of service. Performance has been analysed both with and without congestion. Medium-term congestion for a maximum continuous time span of 10 sec was produced to test traffic isolation between different classes.

The equipment involved in the tests includes as a set of router M160s, three SUN workstations – located in the French, Spanish and Italian PoPs respectively – and two SmartBits 600s by Spirent – located in Frankfurt and Milan. The SmartBits is a network device specialised in traffic generation; the SmartFlows application version 1.50 was adopted to drive the device. Two of the interfaces available on the SmartBits located in the Frankfurt PoP were used: one STM-16 interface connecting it to the M160 DE1.DE.GÉANT.NET and a FastEthernet interface connecting it to DE2.DE.GÉANT.NET. In the latter case connectivity was provided via a Giga/FastEthernet switch connected to router DE2 by a GigaEthernet interface. On the other hand, for the SmartBits in Milan just the STM-16 interface connecting it to the M160 in the Italian PoP was used for traffic generation. The two SmartBits are used to generate large amounts of UDP test traffic (BE, EF and LBE), to produce congestion when needed, and to collect accurate one-way delay measurements thanks to the high accuracy of the SmartBits clock (10 nsec).

While the two SmartBits 600s only managed UDP traffic, the three workstations, running Solaris 2.8, were used to evaluate Best Effort TCP performance with and without LBE congestion. *Netperf* was the main tool used for TCP throughput measurement. All of them are connected to the GÉANT infrastructure by GigaEthernet interfaces.

Figure 2 illustrates the test layout, while Figure 3 provides a general picture of the traffic patterns used during the tests and indicates the location of the two congestion points that were generated when congestion was needed to test traffic isolation.

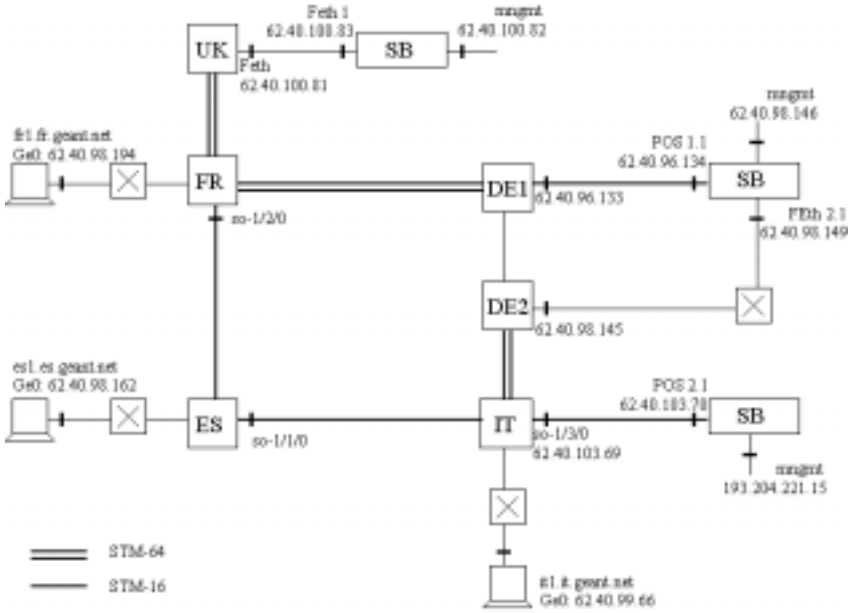

Figure 2: LBE test infrastructure

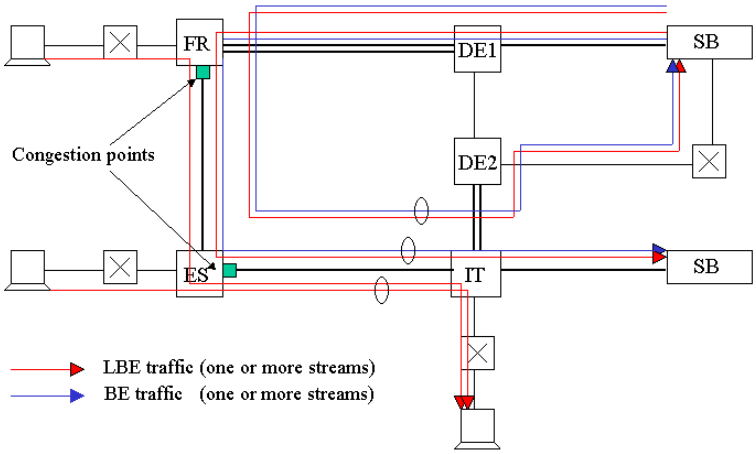

Figure 3: general traffic matrix and the corresponding congestion points

Section 6.2 and 6.3 of this Chapter illustrate the preliminary results of IP Premium, BE and LBE performance without and with congestion respectively.

Unless differently specified, in what follows SmartBits traffic load will be expressed as a percentage of the capacity of the STM-16 interface connecting the SmartBits in Frankfurt – used as main test traffic source – to the network infrastructure.

6.2 LBE and BE performance measurement without congestion

The tests documented in this section are based on the router configuration available in Annex 1. Configurations are explained in Sections 7.3.1 and 7.3.2.

6.2.1 Packet loss and throughput

A variety of LBE traffic loads in the range [10, 20] % for different LBE UDP datagram sizes: {128, 256, 384, 512, 640, 768, 896, 1024, 1152, 1280, 1408} byte were generated by a single constant bit rate LBE stream. Since the background BE production traffic can potentially greatly vary during a test session, the BE traffic volume and BE queues were constantly monitored to make sure that no congestion occurred during the test sessions.

For none of the datagram size/datagram rate combinations mentioned above LBE packet loss was experienced. Even when increasing the LBE traffic load up to 50 % - 1.17 Gbit/sec – with a packet size equal to 60 by no packet loss was observed during test sessions of 10 sec each.

The performance experienced by BE traffic without congestion is identical to the LBE case. In other words, no BE packet loss could be observed. As with LBE traffic, the maximum BE load tested in this case was 50%.

6.2.2 One-way delay

In this test 100 streams are generated by the SmartBits in Frankfurt so that two streams are destined to the FastEthernet interface of the device sourcing traffic, while the remaining 98 streams go to the STM-16 interface of the SmartBits located in Milan. The use of a single device as source and destination at the same time gives the possibility to accurately measure one-way delay, since latency

measures are not affected by clock synchronization errors. One of the flows sourced and received by the same device is BE while the other is LBE. These two reference streams were used for one-way delay measurement and were run concurrently so that a direct performance comparison can be drawn between the two classes. A fraction of the remaining flows received by the SmartBits in Italy is LBE while the remaining part is BE, so that 25 % of the overall traffic load is BE while the remaining part is LBE.

Figure 4 plots the minimum, average and maximum one-way delay experienced by the two reference streams. It can be noticed that in case of no congestion one-way delay is extremely stable: the difference between minimum and maximum is almost negligible for all the traffic loads tested and for both BE and LBE traffic.

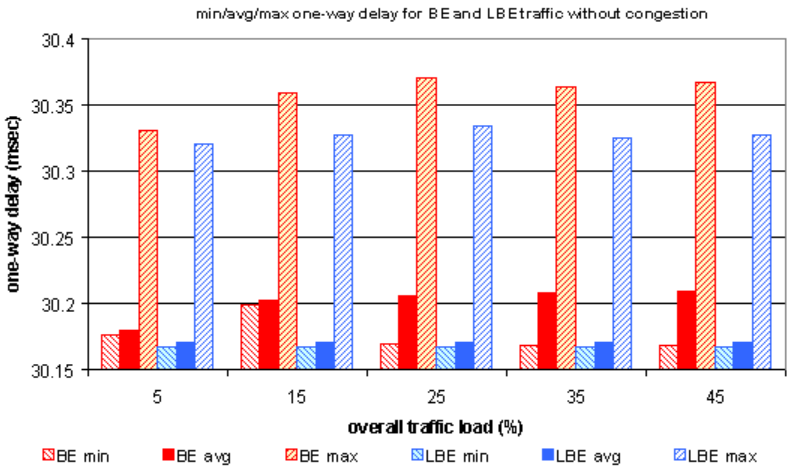

Figure 4: one-way delay for a BE and LBE flow and different traffic loads without congestion

6.2.3 Instantaneous packet delay variation

In case of no congestion, also the instantaneous packet delay variation (IPDV) experienced by one BE stream and one LBE stream, where the two flows are run concurrently and produce one half of the overall load, is comparable for the two classes when the traffic volume varies in the range: [10, 50] %.

Results show that for both services IPDV is almost negligible: the maximum IPDV recorded during the test session was experienced by the BE stream and was approximately equal to only 11 μsec, as shown in Figure 5. In this case, IPDV performance was measured by injecting SmartBits traffic from DE to IT. For a packet sample of 100 consecutive packets, for each traffic class IPDV was computed by calculating the absolute value of the difference of the one-way delay experienced by two consecutive packets. Even if the clocks of the sending and receiving device were not synchronised, the clock offset should not affect the IPDV measurement, since IPDV is a relative metric. The clock skew of the two devices is negligible during each test session, given the short duration equal to 10 sec.

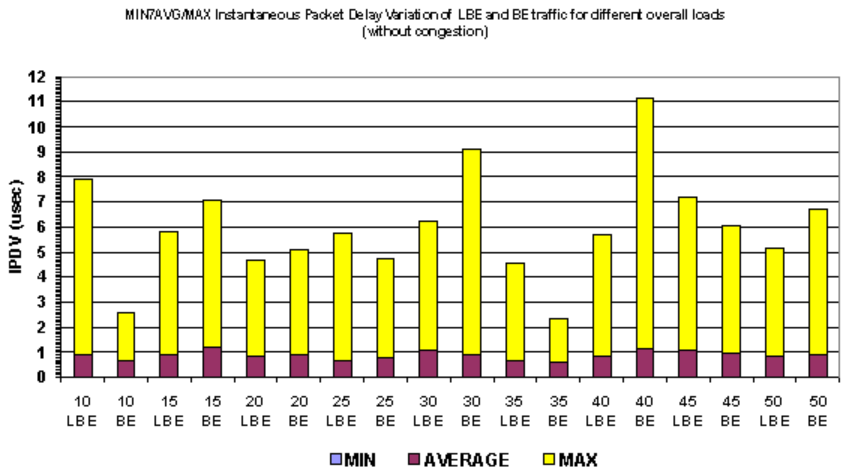

Figure 5: minimum, average and maximum IPDV for BE and LBE traffic and different overall loads. The minimum IPDV is equal to 0 for both BE and LBE for each of the load values

6.2.4 Out-of-sequence packets

A fraction of out-of-order packets was observed. The maximum fraction of out out-of-order packets was experienced with LBE/BE datagrams of 128 by. In our experiments a given packet is counted as out-of-order if its sequence number is *not* equal to one more than the sequence number of the previously received packet. According to this definition, the fraction of out-of-order packets can be accurately measured only in case of no packet loss.

The packet reordering phenomenon also affects the BE stream used for testing, especially for short packet sizes and for large packet rates, as expected. In this test the BE and LBE queue had the same *high* priority as for the LBE test case.

Figure 6 shows that for both BE and LBE traffic the percentage of out-of-order packets is proportional to the packet rate injected. The larger the packet rate, the higher is the probability of receiving some out of sequence packets. As shown later, packet reordering can be greatly reduced through a number of configuration techniques. Reordering is related to the architecture of the router platform under analysis and can be also due to the priority and weight assigned to the LBE and BE queues enabled on a given router interface. In this test both the BE and the LBE queues are assigned the same *high* priority.

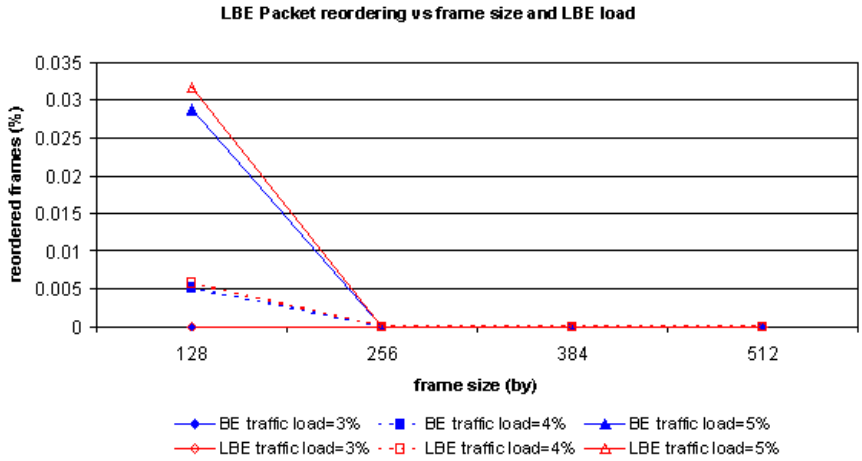

Figure 6: comparison of reordered packets for different BE and LBE flow profiles

6.3 LBE, BE and IP Premium performance in case of congestion

6.3.1 Packet-loss

Independently of the router configurations tested, no packet loss has ever been experienced by BE traffic in case of congestion produced by LBE packets. If both the BE and LBE traffic class are active at the same time and the output interface capacity is exceeded – but the BE offered load is less than the available output capacity – no BE packet loss is ever reported both by the BE queue statistics of the congested interface and by the per-flow packet loss statistics provided by the SmartBits.

For example, when injecting four UDP streams – three LBE and one BE flow – so that the BE offered load is 25 % of the overall test traffic, if the aggregate load varies in the range [95, 96, 97, 98, 99, 100] % of the capacity of a STM-16 line, BE packet loss is always null, while the LBE packet loss percentage is a function of the instantaneous total offered load and in this test can exceed 3 % of the total amount of packets sourced by the SmartBits. The relationship between packet loss and total traffic load for each traffic class is shown in Figure 7.

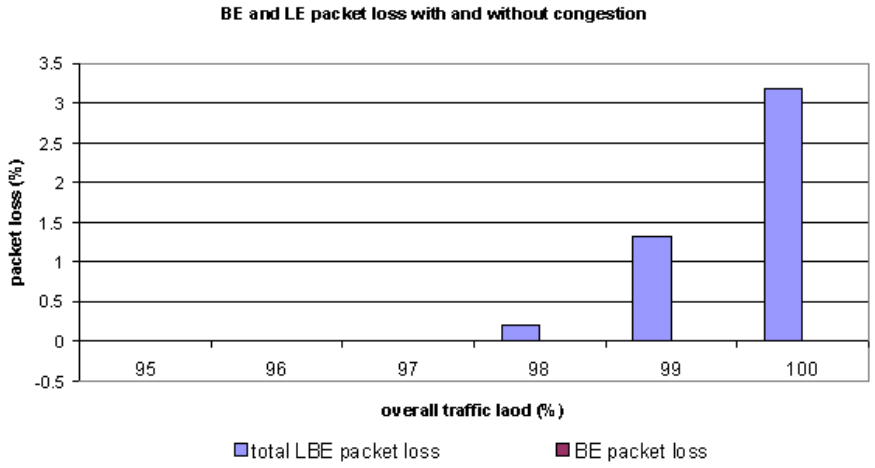

Figure 7: BE and LBE packet loss with and without congestion. No packet loss is experienced by BE traffic for any traffic load

If also IP Premium traffic is present, also the EF class does not experience any packet loss, similarly to what seen for the BE class. No IP Premium and BE loss is present if the aggregate IP Premium and BE load does not exceed the capacity of the output interface.

6.3.2 Throughput

The presence of LBE packet loss is reflected by a decrease of the overall throughput achieved by the LBE streams generated by the SmartBits. As expected, the larger the packet loss rate, the greater the loss in LBE throughput. On the other hand, in case of BE traffic the achieved aggregate throughput equals the traffic rate injected by the SmartBits.

Figure 8 compares the throughput achieved by BE, LBE and IP premium flows injecting traffic at same output rate. The test was run by sourcing 10 streams: seven of them are LBE, three are BE and one is EF. The aggregate load was increased from 50 to 100 %. It can be seen that the throughput of three flows, one of each class of service, is comparable only in case of no congestion. As soon as packet loss occurs (for an overall traffic load equal to 100%), only the LBE stream is affected.

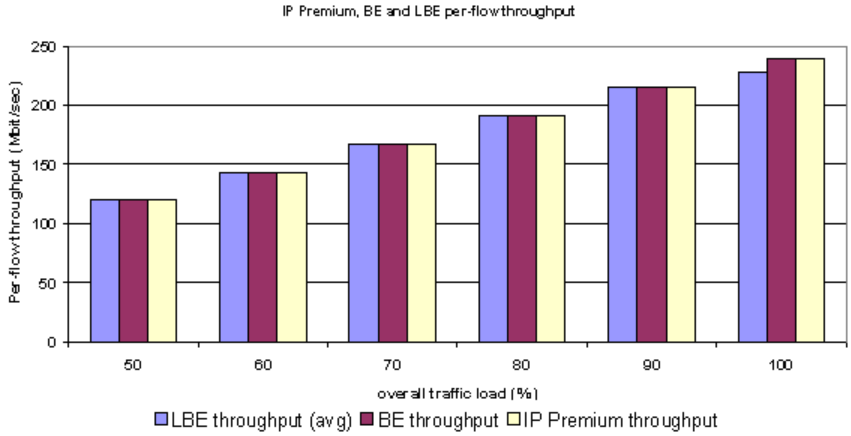

Figure 8: IP Premium, BE and LBE per-flow throughput for different traffic loads. LBE throughput is the average throughput achieved by the seven LBE flows produced by the SmartBits

6.3.3 One-way delay

One-way delay measurement in case of congestion was based on the same traffic pattern described in the previous section for one-way delay measurement in case of no congestion, i.e. with 100 streams of which 2 reference streams (one BE and one LBE) are sourced and received by the same SmartBits. In this case the overall amount of test traffic produced by the SmartBits is higher and varies in the range: [95, 100] %.

While no effect on one-way delay could be observed in case of no congestion for different traffic loads, a different behaviour is shown in case of congestion.
As indicated in Figure 9, LBE traffic experiences an increase of both average and maximum one-way delay when congestion starts, while the minimum latency is constant. Also BE traffic experiences a slight increase in one-way delay, but in case of BE traffic the increase is negligible. The maximum difference between minimum and maximum one-way latency for LBE traffic – experienced with 100 % of overall traffic – is 1.865 msec, while the maximum difference for LBE traffic, which was observed in similar traffic load conditions, is only 0.537 msec.

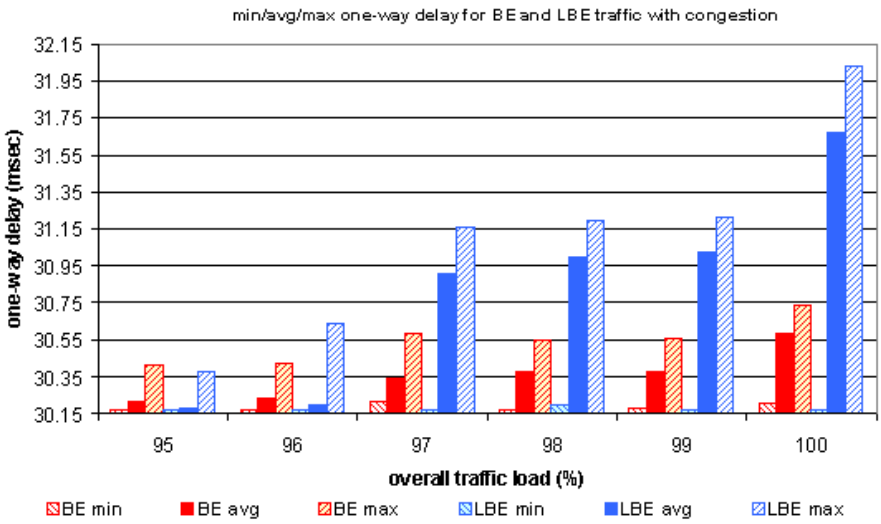

Figure 9: minimum, average and maximum one-way delay for a BE and LBE flow and different traffic loads with and without congestion

The analysis of one-way delay for both LBE and BE traffic with and without congestion through the calculation of one-way frequency distributions gives the possibility to better understand the influence of congestion on the entire packet population.

For this test frequency distributions are calculated according to the following method. Time is divided in intervals of 0.1 sec each. For each interval the minimum, average and maximum one-way delay are recorded. The overall test length is 10 sec so that 100 samples are generated.

As illustrated in Figure 10 and 11, both in case of LBE and BE traffic congestion produces a shift of the frequency distribution peak to the right and a change in the curve profile, in fact, in both cases the distribution tails get longer. The different length of the tail in the two cases can be noticed by comparing the horizontal axis scales of the graphs of Figures 10 and 11. We see that the shift of the BE distribution is much more limited than for LBE traffic. In general, also in this case we can conclude that the impact of congestion on one-way delay profile of BE traffic is almost negligible.

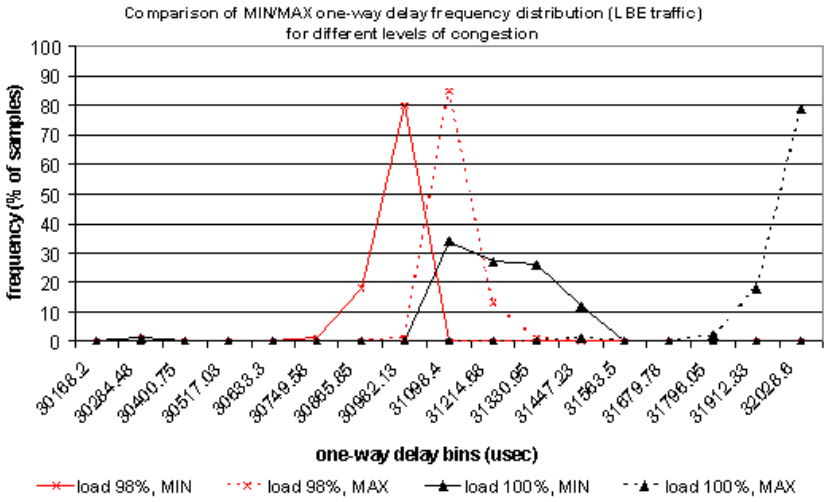

Figure 10: comparison of *LBE* frequency distributions of the minimum and maximum one-way delay without congestion (when load is equal to 98%) and with congestion (when load is equal to100%).

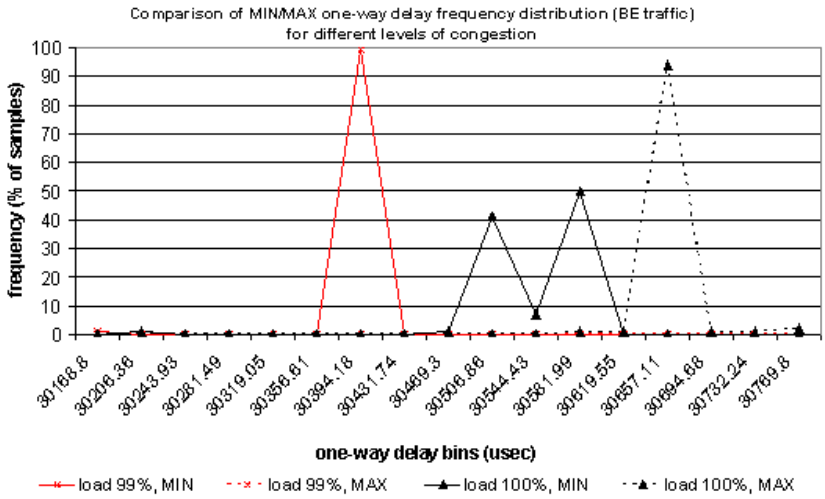

Figure 11: comparison of *BE* frequency distributions of the minimum and maximum one-way delay without congestion (when load is equal to 98%) and with congestion (when load is equal to 100%)

6.3.4 Instantaneous packet delay variation

IPDV was tested with the same traffic profile used in case of no congestion, i.e. with two individual flows: a BE and a LBE flow generated by the SmartBits in Germany and received by the SmartBits in Italy. In this case, the overall traffic load is higher and varies in the range: [95, 100] %.

Figure 12 shows that the maximum LBE IPDV tends to increase with congestion, i.e. for high traffic loads, while the minimum and average IPDV are not dependent on either traffic load or congestion. Figure 13 plots the IPDV frequency distribution computed over a sample of 100 consecutive LBE packets and shows that the two distributions with congestion (100% of the capacity of a STM-16 interface) and without congestion (95% of the same capacity) are equivalent.

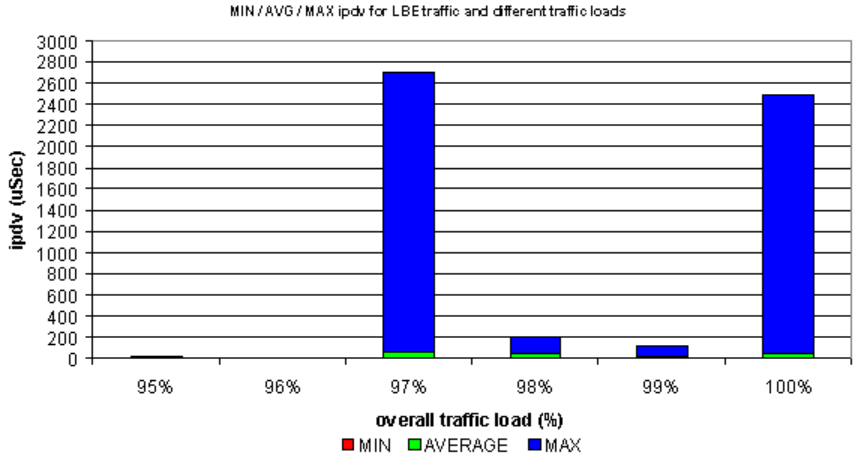

Figure 12: minimum / average and maximum IPDV experienced by LBE traffic in presence of different traffic loads

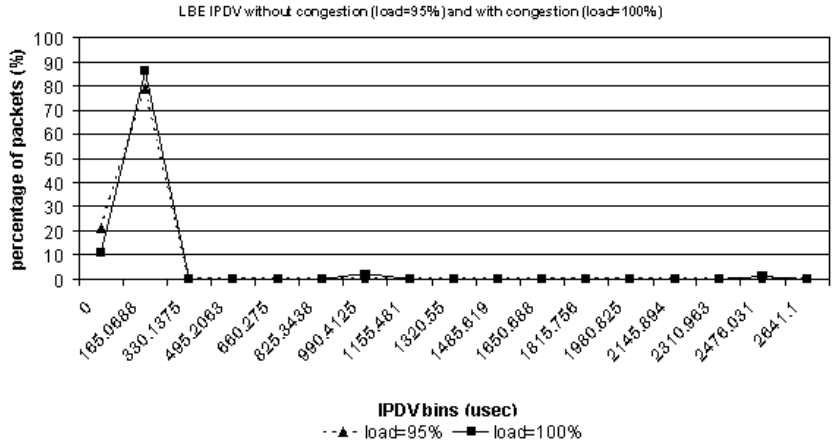

Figure 13: comparison of IPDV frequency distributions for LBE traffic with and without congestion

6.3.5 Out-of-sequence packets

As already reported in the previous section, out-of-sequence packets can be observed for flows of different traffic classes for a large range of flow rates. Test results show that the packet-reordering behaviour can be greatly influenced by the type of *priority* assigned to the BE and LBE queues.

In the first case, when the LBE priority is equal to the BE queue priority, see section 7.3.1 for more information about the configuration, the percentage of out-of-order packets tends to increase exponentially in case of congestion and is a function of the BE flow packet rate. Figure 14 compares the percentage of out-of-order packets for the same overall BE and LBE traffic loads but for BE flows

injecting traffic at *different* rates. It can be seen that while for low-rate flows a negligible percentage is observable only in case of congestion.

For higher-rate flows out-of-order packets are already experienced without congestion. The percentage can be more than 80 % when congestion appears.

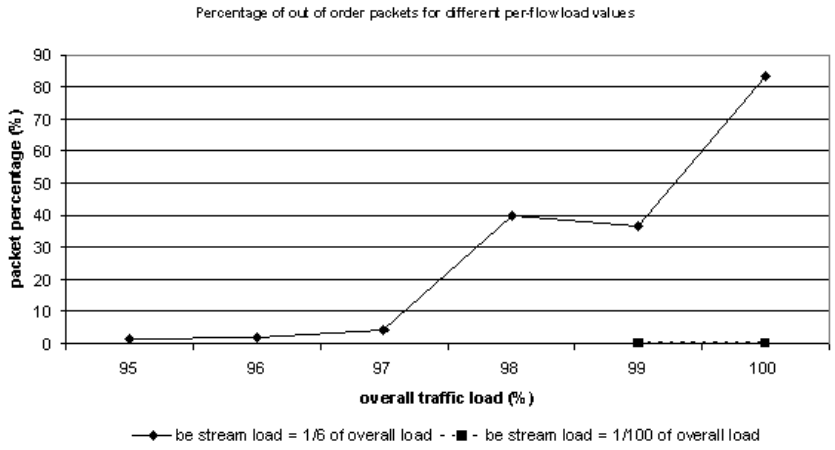

Figure 14: Percentage of out of order BE packets for different overall traffic loads and different per-flow loads

Results also show that the presence of out-of-order packets for a given flow is a function of the amount of background traffic present in the traffic class of the flow, and increases in case of congestion.

In case of different queue priorities, (see section 7.3.2 for more information), tests show that the percentage of out-of-order packets increases linearly with the amount of traffic load produced for that class, as illustrated in Figure 15. Note that in this test the BE streams used for performance comparison had the same output rate, independently of the overall BE traffic load. In fact, the overall number of BE and LBE flows was always the same and equal to 10 for each traffic load. The increase in BE load was simply generated by adding BE streams and subtracting LBE streams.

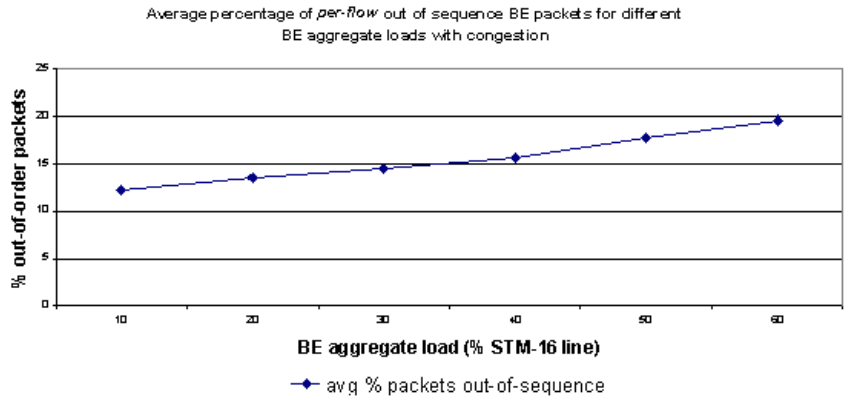

Figure 15: average percentage of per-flow BE out-of-order packets as a function of the overall amount of test BE traffic.

While even in case of different queue priorities BE and LBE packet reordering is not completely avoided, the aggregate *Best Effort* TCP load observed by reference flows generated by test workstations in the PoPs is constant and independent of the amount of background BE test traffic.

Obviously, this is provided that BE traffic does not exceed the capacity of any of the output interfaces on the data path.

In other words, in case of different queue priorities, the packet reordering observed does not have a negative effect on the end-to-end TCP performance, as shown in Figure 16.

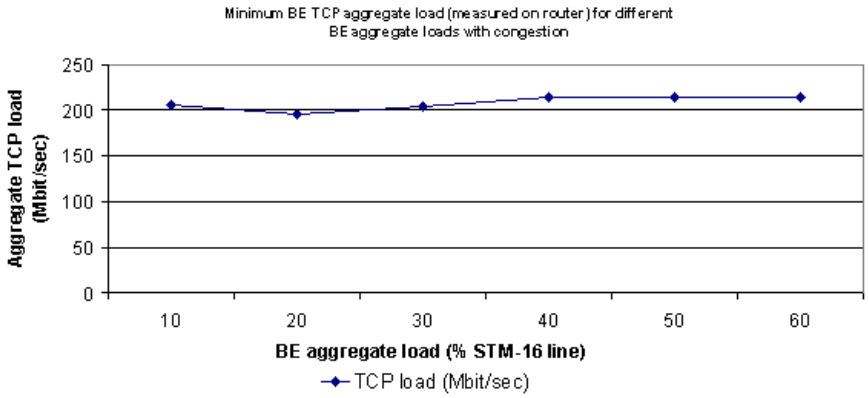

Figure 16: aggregate BE TCP load reported by the M160 as a function of the aggregate BE load produced by the SmartBits

6.3.6 Conclusions

The preliminary test results of the LBE service are extremely promising. They show that the congestion produced by large volumes of LBE traffic has no effect on BE and IP Premium performance in the traffic scenarios used during the test sessions. In particular, if the amount of available bandwidth in a given router is not exceeded by BE and IP Premium taffic, no packet loss for these two classes is introduced by the addition of LBE traffic. Moreover, congestion seems to cause an increase in BE average one-way delay that we think is almost negligible. The one-way delay and IPDV profile of LBE traffic itself is not particularly affected by congestion either: under the most severe congestion conditions the average one-way delay of LBE packets only increase by less than 1 msec, while the average LBE IPDV does not seem to increase at all.

We have to note that congestion increases the percentage of out-of-order packets received by all traffic classes. However, by appropriately configuring queue priorities packet-reordering effects can be controlled so that no adverse impact on TCP end-to-end performance can be observed.

7 GÉANT ROUTER CONFIGURATION

This section explains the configuration used on the GÉANT's Juniper M-160[2] routers for the Less than Best Effort tests. It also provides explanation about the configuration choices.

The LBE service can be characterised by the following property: "a very small percentage of network capacity is allocated to LBE so that, under congestion, the BE traffic and any higher priority traffic classes are protected from LBE traffic". The configuration applied on the GÉANT routers have thus to be engineered in order to support this. The Best Effort and the Premium IP are the higher priority classes of interest in the GÉANT network and they can be defined as follows:

- *Best Effort*: default class of service offered by a network. No guarantees are provided for the best effort traffic, except that the network will do its best to transmit the packet.;

- *Premium IP*: class of service for which the packets have an upper bounded one-way delay (OWD), an upper bounded IP Packet Delay Variation (IPDV) and upper bounded one-way packet loss (OWPL).

The traffic from these services can potentially starve LBE packets. In fact, the LBE can only use the network capacity which has not been used by the Best Effort and Premium IP classes. In any case, the profile of both Premium IP flows (OWD, OWPL and IPDV) and of BE flows should not be distorted meaningfully by the presence of LBE packets.

Some preliminary additional tests similar to the ones done on GÉANT, have been done in laboratory; the GÉANT configurations applied on the production infrastructure have been tested. The goal of these tests was to verify the best effort and less than best effort services behaviour with very high BE traffic load[3].

7.1 The M-series queuing architecture

An M-series Juniper has four output queues per port[4]. On GÉANT, they are named BE, LBE, Premium and Network Control. A Weighted Round Robin (WRR) scheduler algorithm serves these output queues on the basis of two parameters per queue: the queue weight and the queue priority.

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;

[2] This configuration can be applied to any M-series Juniper router based on E-FPCs (Enhanced-FPC). On the other hand, most of the results obtained on a M-160 can be extended to the whole M-series range of products. The difference is due to the different architectures available on the M-160 and on other M-series routers.
[3] These tests could not have been done on GÉANT for the safety of operational traffic.
[4] The interested reader is encouraged to refer to Juniper documentation for more information on E-FPCs.

4. Low priority, negative credit queues.

The following explanation tries to clarify the WRR mechanism[5].

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 dedicated bandwidth and no more packers are present in a "positive credited" queue.

The leftover bandwidth (from the positive credited queues) is fairly shared[6] 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.

Packets are mapped to output-queues according to their DSCP tagging:
- DSCP 46 (Premium IP) into the Premium queue;
- DSCP 8 (LBE) into the LBE queue;
- DSCP 48 and 56 (NC) in the Network Control queue;
- Any other DSCP is mapped to the BE queue.

7.2 GÉANT configuration before the tests

The configuration on the GÉANT routers was set-up to provide BE and Premium IP services.

Queue	FC[7]	Service	DSCP	Weight	Priority
0	BE	Best Effort	0	5%	Low
1	EF	Premium IP	46	90%	Low
2	-	-	-	-	-
3	NC	Network control	48/56	5%	Low

Table 1 – Current GÉANT queuing scheme.

Thanks to their weights, the Premium IP and Network control queues have a reserved bandwidth higher than their utilisation level. Their leftover bandwidth is fully allocated to the Best Effort traffic.

Queue 2 was chosen to host the Less than Best Effort traffic since it is currently unused.

7.3 Test configurations

The various configurations adopted on the routers are described in this section. Configurations have been gradually refined according to the increasing detailed understanding of the WRR algorithm acquired during the tests.

[5] The explanation does not attempt to provide a complete picture of the queuing behaviour on a per-packet basis, but it rather gives a general picture of it.
[6] This sharing is done independently of the queue weight.
[7] In Juniper devices Forwarding Class (FC) identifies the queue name.

7.3.1 First configuration

A weight of one percent is allocated to LBE queue (what was thought as being the minimum configurable weight).

Queue	FC	Service	DSCP	Weight	Priority
0	BE	Best Effort	0	5%	Low
1	EF	Premium IP	46	90%	Low
2	LBE	Less than BE	8	**1%**	Low
3	NC	Network control	48/56	4%	Low

Table 2 – First choice of weight allocation for the LBE queue.

During the test described in Section 6.3, when BE and LBE had the same priority, it was noticed that the Best Effort traffic was not protected enough in terms of end-to-end TCP throughput, which tended to decrease in case of congestion when a relatively high BE load is produced by the SmartBits. According to our expectations, non-significant drop of BE TCP throughput should have been experienced, since the SmartBits reported no BE packet loss.

One possible explanation of the throughput drop was a too small difference in weight between the BE and the less than best effort queues[8]. For this reason, the LBE weight was then reduced to zero[9].

Queue	FC	Service	DSCP	Weight	Priority
0	BE	Best Effort	0	5%	Low
1	EF	Premium IP	46	90%	Low
2	LBE	Less than BE	8	**0%**	Low
3	NC	Network control	48/56	5%	Low

Table 3 – A weight of zero is allocated to the LBE queue.

Unfortunately, this configuration change didn't solve the BE TCP throughput loss problem.

Once the Premium is served[10] the remaining bandwidth unused by Premium traffic[11] is allocated to the BE and the LBE queues. These queues quickly get a negative credit since the amount of offered load usually exceeds the bandwidth allocated to them. The bandwidth is equally shared between negative credit queues with the same priority. Additional lab tests showed that, with this configuration, if the BE offered load is too high, the BE class could suffer from losses instead of the LBE one. This is clearly non-LBE compliant.

There are two potential ways of solving the problem:
- The increase of the weight allocated to the BE queue;

[8] The parameter that is actually responsible of such throughput decrease is still unknown.
[9] A weight of zero doesn't mean that the Juniper's Weighted Round Robin (WRR) scheduler never services the queue. The WRR visits every queue, even the ones for which no weight is configured (no weight means either a queue not configured or a queue a weight of zero is allocated to).
By default, the WRR serves one byte per round out of the "zero-weighted queues". The service rate of these queues can increase if some bandwidth is left by the "non-zero weighted" queues, i.e. when "non-zero weighted queues" are empty.
[10] The Premium IP bandwidth utilisation should never be higher than 10% of the link capacity.
[11] The Network Control traffic volume is very low too and does not use its entire credit. The bandwidth unused by the NC traffic is also redistributed to the BE and LBE queue. However, given the small weight assigned to the NC queue, the amount of NC spare capacity is much lower than for IP Premium.

- A different setting of queue priorities so that Premium, BE and NC queues have *high* priority while the LBE priority is *low*.

7.3.2 Final configuration

In order to avoid a significant decrease of the weight assigned to the IP Premium queue, the second solution was adopted, so that in the third configuration tested the Priority queue statement is added.

Queue	FC	Service	DSCP	Weight	Priority
0	BE	Best Effort	0	5%	**High**
1	EF	Premium IP	46	90%	**High**
2	LBE	Less than BE	8	**0%**[12]	Low
3	NC	Network control	48/56	5%	**High**

Table 4 – A weight of zero and a low priority is allocated to the LBE queue. Other queues are assigned a high priority.

In this way, when both BE and LBE queues have a negative credit, it is always the BE queue which is served first until it's empty. Only at this point is the LBE queue served. With this configuration, no significant drop of BE TCP throughput was observed during congestion.

[12] 0% was configured for the LBE queue in the previous phase and was kept in the final configuration. A weight of 1% could have been chosen and the results would have been similar. The most important point of the last configuration is the "low" priority attributed to the LBE queue while the other queues have a "high" priority. For a better understanding of the difference between 0% and 1% for the LBE queue, see section 7.1.

8 CONCLUSIONS AND FUTURE WORK

In this report we have proposed a service description for a new LBE service whose primary function is to be able to make use of available bandwidth, but in such a way as to always defer to BE (or better) traffic where congestion occurs (LBE packets are always dropped before BE packets). We have run a first set of tests to evaluate this proposed LBE service. The results presented in Chapter 6 are encouraging in that they largely validate the feasibility of operating an LBE service on the GÉANT backbone. The LBE traffic in the configuration tested does not adversely affect the regular BE or Premium IP traffic. As concluded in Section 7.3.2, no significant drop of BE TCP throughput was observed during congestion.

One of the concerns that arose during testing lay in the high observed rate of (BE) packet reordering that was occurring under congestion in the presence of LBE traffic. However, test results show that the packet-reordering behaviour can be greatly influenced by a number of configuration techniques including the type of *priority* assigned to the BE and LBE queues, as reported in Section 6.3.5.

The tests performed to date focus on the network backbone (GÉANT). The results are also applicable to NREN networks. However, we note that in the context of GÉANT most network congestion is occurring at the edges of the network; thus implementation at the edge (university access links) is also important, especially where a university may be receiving (or sending) LBE traffic and giving it equal treatment to BE traffic at the campus edge router

It is also important to note that it is the sender of the data that must mark the traffic with the LBE DSCP. Thus a user who wishes to initiate an LBE FTP session requires a way to signal this request to the FTP server. An example solution for such a requirement is described in Section 3, but the general question of LBE signalling, and of voluntary against enforced use of LBE, is open for further investigation.

Having identified a number of scenarios to which LBE can be applied in Section 3, the logical next step is to progress with the implementation of LBE on the GÉANT backbone, to promote its adoption within the NRENs (or at least DSCP transparency - the routers do not reset the LBE DSCP when observed) and to then encourage end users to start using the service.

To help in this process, we will also shortly run a further set of tests as described in Section 5, where we evaluate the use of LBE where the intermediate network may offer only DCSP transparency, and the queuing and drop policy is applied at the network edge (the university) where congestion occurs. The results of these tests will be presented as an Addendum to this report.

This report describes the Scavenger service running on Internet2. We have shown how the LBE service presented here is interoperable with Scavenger by use of a common DSCP value. We also plan to identify LBE trial applications for use with other networks (e.g. in Japan) and to promote an interoperable service in those networks.

Finally, we observe that the implementation of LBE is IP-independent. This report only considers IPv4 networks. We hope to be able to run an LBE service over IPv6 within the 6NET project (the partners identified for the further tests in Section 5 are both 6NET partners).

9 ACKNOWLEDGMENTS

We thank Spirent for having given us the possibility to use SmartBits 600s on a loan basis , the SmartFlow 1.50 software and for their support. The accuracy in latency, packet loss and packet reordering measurement has been extremely important to understand the dynamics of the LBE service and its interaction with Best Effort and IP Premium traffic in the GÉANT infrastructure. We also thank our colleagues from the DataGrid [EDG] Work Package 7 for the collaboration provided for the definition of the LBE test programme and the participation to the tests documented in Section 6. Such collaboration is in the framework of a co-operation agreement established in June 2001and of a Memorandum of Understanding defined between the DataGrid and GEANT projects."

10 REFERENCES

[6net] The 6NET Project
http://www.6net.org/

[abe] Alternative Best Effort (ABE)
http://www.abeservice.com

[abilene] The Abilene network
http://www.ucaid.org/abilene/

[bless-le] A Lower Effort Per-Domain Behavior for Differentiated Services, Internet Draft, June 2002, R. Bless et al,
http://www.ietf.org/internet-drafts/draft-bless-diffserv-pdb-le-00.txt (temporary)

[chariot] Chariot Performance Measurement Suite
http://www.netiq.com/products/chr/default.asp

[diffserv] IETF diffserv Working Group
http://www.ietf.org/html.charters/diffserv-charter.html

[e2epi] Internet2 End-to-End Performance Initiative
http://e2epi.internet2.edu/index.shtml

[EDG7.3] Network services: requirements, deployment and use in testbeds, DataGrid project deliverable DataGrid-07-D7-3-0113, June 2002

[EDG] The DataGrid project, http://eu-datagrid.web.cern.ch/eu-datagrid/Intranet_Home.htm

[geant-d91] Specification and Implementation Plan for a Premium IP service
http://www.dante.net/tf-ngn/GEA-01-032.pdf
http://www.dante.net/tf-ngn/GEA-01-032av2.pdf (Implementation addendum)
http://www.dante.net/tf-ngn/GEA-01-032b.pdf (SLA addendum)

[geant-d94] GÉANT Deliverable D9.4: Testing of Traffic Measurement Tools
http://www.dante.net/tf-ngn/D9.4v2.pdf

[i2-prem] QBone Premium IP service (and deployment problems thereof)
http://qbone.internet2.edu/premium/

[i2-qoswg] Internet2 QoS Working Group
http://www.internet2.edu/qos/wg/

[prem-ip] Premium IP service
http://axgarr.dir.garr.it/~cmp/tf-ngn/IPPremium.html

[qbss] QBone Scavenger Service
http://qbone.internet2.edu/qbss/

[qbone] QBone Initiative
http://qbone.internet2.edu/

[RC] Regional Centers for LHC computing; the MONARC Architecture Group, http://monarc.web.cern.ch/MONARC/docs/monarc_docs/1999-03.html

[tfngn-lbe] TF-NGN LBE WG http://www.cnaf.infn.it/~ferrari/tfngn/lbe/

11 ACRONYMS

BE	Best Effort
Diffserv	Differentiated Service
E-FPC	Enhanced – Flexible PIC Concentrator
EF	Expedited Forwarding
FC	Forwarding Class
FPC	Flexible PIC Concentrator
IP	Internet Protocol
IPDV	IP Packet Delay Variation
LBE	Less than Best Effort
NC	Network Control
OWD	One Way Delay
OWPL	One Way losses
PIC	Physical Interface Card
QBSS	QBone Scavenger Service
QoS	Quality of Service
WFQ	Weighted Fair Queuing
WRR	Weighted Round Robin

A1. ANNEX 1 – ROUTER CONFIGURATION

The configuration shown applies to any JUNOS version starting from JUNOS 5.1 and to M-series routers using E-FPC (Enhanced Flexible PIC Concentrators).

A.1 Last (third) GÉANT router configuration

The configuration showed below is the command corresponding to the last of the three configurations described previously in section 7. It correspond to the one with the high priority for the best effort, expedited forwarding and network control queue and low priority for the less than best effort one. See also section 7.3.2.

```
class-of-service {
    classifiers {
       dscp lbe-classifier {
            import default;
            forwarding-class best-effort {
                loss-priority low code-points [ af11 af12 af13 ];
            }
            forwarding-class less-than-best-effort {
                loss-priority low code-points cs1;
            }
        }
    }
    drop-profiles {
        best-effort-drop-profile {
            fill-level 15 drop-probability 30;
            fill-level 19 drop-probability 50;
            fill-level 24 drop-probability 75;
            fill-level 30 drop-probability 100;
        }
        expedited-forwarding-drop-profile {
            fill-level 10 drop-probability 100;
        }
        default-drop-profile {
            fill-level 100 drop-probability 100;
        }
    }
    forwarding-classes {
        queue 0 best-effort;
        queue 1 expedited-forwarding;
        queue 2 less-than-best-effort;
        queue 3 network-control;
    }
    interfaces {
        so-1/2/0 {
            scheduler-map MAP-BACKBONE-2;
            unit 0 {
                classifiers {
                    dscp lbe-classifier;
                }
                rewrite-rules {
                    dscp dscp-lbe-2;
                }
            }
        }
        so-1/2/0 {
            scheduler-map MAP-BACKBONE-2;
            unit 0 {
                classifiers {
                    dscp lbe-classifier;
                }
                rewrite-rules {
                    dscp dscp-lbe-2;
                }
```

```
            }
        }
    }
    rewrite-rules {
       dscp dscp-lbe-2 {
            import default;
            forwarding-class best-effort {
                loss-priority low code-point be;
            }
            forwarding-class less-than-best-effort {
                loss-priority low code-point cs1;
            }
        }
    }
    scheduler-maps {
        MAP-BACKBONE-2 {
            forwarding-class best-effort scheduler sch-backbone-best-effort;
            forwarding-class expedited-forwarding scheduler sch-backbone-expedited-forwarding;
            forwarding-class network-control scheduler sch-backbone-network-ctrl-2;
            forwarding-class less-than-best-effort scheduler sch-less-than-best-effort;
        }
    schedulers {
        sch-backbone-best-effort {
            transmit-rate percent 5;
            buffer-size percent 5;
            priority high;
        }
        sch-backbone-expedited-forwarding {
            transmit-rate percent 90;
            buffer-size percent 90;
            priority high;
        }
        sch-backbone-network-ctrl {
            transmit-rate percent 5;
            buffer-size percent 5;
            priority high;
        }
        sch-less-than-best-effort {
            transmit-rate percent 0;
            buffer-size percent 0;
            priority low;
        }
    }
}
```

A.2 First GÉANT router configuration

This configuration is the first one that was used on the GÉANT router. A weight of one was allocated to the less than best effort queue. This section highlights the differences with the section 11.1. See also section 7.3.1.

```
class-of-service {
    scheduler-maps {
        MAP-BACKBONE-2 {
            forwarding-class best-effort scheduler sch-backbone-best-effort;
            forwarding-class expedited-forwarding scheduler sch-backbone-expedited-forwarding;
            forwarding-class network-control scheduler sch-backbone-network-ctrl-2;
            forwarding-class less-than-best-effort scheduler sch-less-than-best-effort;
        }
    schedulers {
        sch-backbone-best-effort {
            transmit-rate percent 5;
            buffer-size percent 5;
        }
        sch-backbone-expedited-forwarding {
            transmit-rate percent 90;
            buffer-size percent 90;
        }
        sch-backbone-network-ctrl {
            transmit-rate percent 4;
            buffer-size percent 4;
        }
        sch-less-than-best-effort {
```

```
            transmit-rate percent 1;
            buffer-size percent 1;
        }
    }
}
```

A.3 Second GÉANT router configuration

The configuration is quite similar to the previous one, except for the weight of the network-control and LBE forwarding-class. See also section 7.3.2.

```
class-of-service {
    scheduler-maps {
        MAP-BACKBONE-2 {
            forwarding-class best-effort scheduler sch-backbone-best-effort;
            forwarding-class expedited-forwarding scheduler sch-backbone-expedited-forwarding;
            forwarding-class network-control scheduler sch-backbone-network-ctrl-2;
            forwarding-class less-than-best-effort scheduler sch-less-than-best-effort;
        }
    schedulers {
        sch-backbone-best-effort {
            transmit-rate percent 5;
            buffer-size percent 5;
        }
        sch-backbone-expedited-forwarding {
            transmit-rate percent 90;
            buffer-size percent 90;
        }
        sch-backbone-network-ctrl {
            transmit-rate percent 5;
            buffer-size percent 5;
        }
        sch-less-than-best-effort {
            transmit-rate percent 0;
            buffer-size percent 0;
        }
    }
}
```