Extending QoS support from Layer 3 to Layer 2

Christos Bouras Vaggelis Kapoulas Vassilis Papapanagiotou Leonidas Poulopoulos

Dimitris Primpas Kostas Stamos

Research Academic Computer Technology Institute, Greece
and Computer Engineering and Informatics Dept., University of Patras, Greece
Email: {bouras,kapoulas,papapana,leopoul,primpas,stamos}@cti.gr

***Abstract*—This paper presents some of the results obtained by the application of Ethernet Layer 2 Quality of Service in IP networks. IP networks traditionally provide Quality of Service in Layer 3. However, since there is an enormous existing Layer 2 infrastructure, today's networks could benefit from the deployment of Layer 2 Quality of Service and the cooperation between Layer 2 Quality of Service and Layer 3 Quality of Service. In this paper, experiments are suggested and conducted and a scheme is suggested for efficient cooperation between Layer 2 and Layer 3 QoS provisioning.**

I. Introduction

Quality of Service (QoS) provisioning has become indispensable in today's networks. Most existing QoS solutions are deployed in Layer 3 (network layer). In order to provide end-to-end QoS guarantees in these networks, the need for Layer 2 QoS deployment as well as the cooperation between any existing Layer 3 QoS deployment must be studied.

QoS provisioning in Layer 2 is very important to networks that are primarily based on Layer 2 infrastructure as it is the only way to provide QoS on the network. Furthermore, networks based on both Layer 2 and Layer 3 network devices could benefit from a more integrated approach in end-to-end QoS provisioning that includes both Layer 2 and Layer 3. Moreover, Layer 2 QoS is lightweight, easily implemented and independent of Layer 3. Because its independency, it can also be applied to non-IP networks where any QoS provisioning was impossible or very difficult. In this paper, we examine the cooperation between Layer 2 and Layer 3 QoS in IP networks. When discussing Layer 2 devices and procedures in this paper, we are specifically referring to Ethernet technology switches, which have become the dominating Layer 2 technology during the past years and have largely substituted older technologies at the same layer, such as ATM and Frame Relay.

Layer 2 Ethernet switches rely on 802.1p standard to provide QoS. The standard 802.1p is part of the IEEE 802.1Q [5] which defines the architecture of virtual bridged LANs (VLANs). This architecture uses tagged frames inserted in Ethernet frames after the source address field. One of the tag fields, the Tag Control Information, is used by 802.1p in order to differentiate between the classes of service. More specifically, the 3 most significant bits of the Tag Control Information field known as Priority Code Point (PCP) are used to define frame priority. Taking advantage of PCP, QoS in Layer 2 can be applied.

An overview of Ethernet L2 QoS capabilities has been given in [9], while Layer 2 QoS experiments with Ethernet switches have been conducted and described in [1]. In [1] 4 Layer 2 QoS experiments are conducted and effects on link throughput and packet loss are shown. Other researchers such as [8] have dealt with Layer 2 QoS in ATM networks. An interesting application of L2 Ethernet QoS has been studied in the field of avionics networks with the demand for low latency and jitter in [10] and [11], while 802.1p has been studied as an approach for the improvement of traffic performance originating from collaborative systems applications in [12].

In this paper, we present how it is possible to deploy Layer 2 QoS in the Greek Research Network where Layer 3 has already been deployed, and how we can make them cooperate. Three different experiments are suggested and implemented. The experiments were conducted in a laboratory environment with the aim to simulate and approximate as close as possible the Greek Research Network production environment and in order to properly design and implement the service in a large scale. Effects of the application of Layer 2 QoS on throughput, packet loss and jitter are discussed and explained. In addition, limitations of the network devices are mentioned along with the problems that were faced and their workarounds.

This paper is organized as follows: Section 2 gives a description of the problem. Section 3 discusses implementation issues, test environment and tools that were used. Section 4 summarizes the experiments conducted and discusses their results. Furthermore, Section 5 examines how to handle QoS deployment when having multiple Layer 2 paths as in Crete's MAN in Greece. Finally, in Section 6 the conclusions and recommendations for future work are presented.

II. Service Architecture

Grnet is the Greek National Research and Education Network (NREN) [2]. Grnet is a mixed IP- and Ethernet-based network, operating at Gigabit speeds. Together with the high-speed LANs of its subscribers (universities and research institutes) and the European academic and research backbone,

978-1-4244-2036-0/08/$25.00 ©2008 IEEE

GEANT, Grnet forms a set of hierarchically-federated networks.

The Grnet backbone consists of network nodes in 8 major Greek cities, namely, Athens (3 PoPs), Thessaloniki, Patras, Ioannina, Xanthi, Heraklion, Larisa and Syros. The WAN network is built on DWDM links with 2.5Gbps capacity (STM-16 lambdas). The access interfaces of the routers are using Gigabit Ethernet technology and connect the 70 subscribers of Grnet (universities, research institutes and the school network). In addition to the WAN, Grnet also contains 2 distinct MAN networks. The Athens MAN is router-based (Figure 1), whereas the Crete MAN is switch based (Figure 2), with a router in the main aggregation site (Heraklio). Both networks are built on unprotected DWDM rings; the Athens MAN uses STM-16 lambdas, whereas the Crete MAN operates on 1-Gigabit Ethernet lambdas.

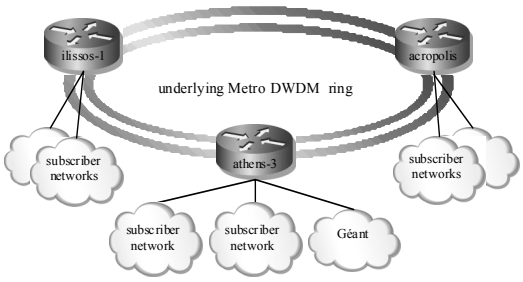

Figure 1: L3 Athens MAN

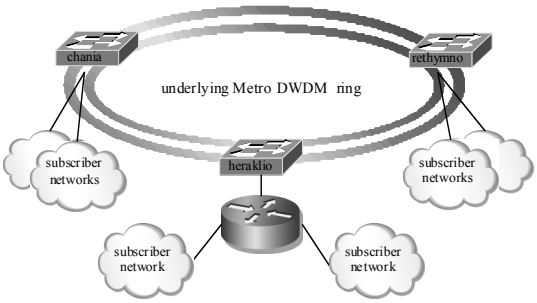

Figure 2: L2 Crete MAN

The Greek Research and Academic Network (GRNET [2]) has deployed for several years a Layer 3 QoS service based upon the features provided by the MPLS technology deployed in the core of the GRNET network, and DiffServ architecture. This architecture allows the support of multiple classes of service. The focus is on three separate classes of service, namely IP Premium for absolute performance guarantees, best effort for the usual treatment of traffic packets and Less than Best Effort (LBE) for non-critical traffic that can be dropped first in case of congestion. IP Premium service is a circuit-like subscriber-to-subscriber service, where both subscriber end-networks and the necessary bandwidth allocation are known at request time. IP Premium service is provided using a provisioning tool called ANStool [4][14]. LBE is provided unprovisioned, which means that each subscriber decides on its own and uses this service simply by marking the packets appropriately. In order to provide the QoS service, the Layer 3 network equipment (routers) has to perform traffic marking, classification, policing and shaping. Per-flow functions are performed at the edge routers of GRNET network, while core routers only perform per-traffic class functions, based on the MPLS Exp field.

The above service design has several implications for traffic between two GRNET clients (such as institutions, universities or other research organizations). It means that traffic coming out of GRNET network ("output" for GRNET edge routers) has been subjected to the specified QoS mechanisms. However, traffic coming into the GRNET network ("input" for GRNET edge routers) receives no treatment up to the point of reaching the edge Layer 3 device (router) of the GRNET network.

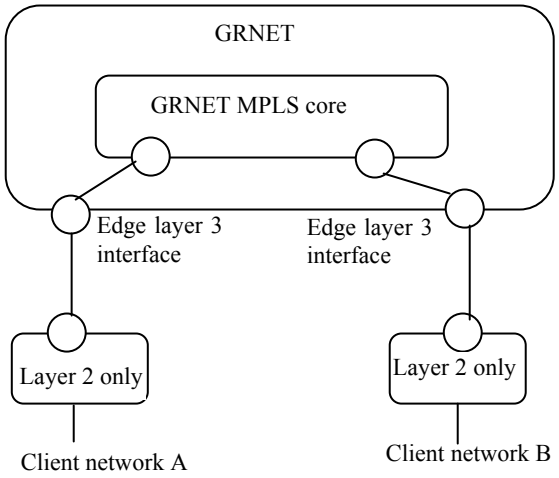

Figure 3: Schematic of GRNET core/edge/L2-only edge network parts

In the most common case (except Crete's MAN), traffic between the GRNET client and the GRNET edge router will go through one or more Layer 2 devices (Ethernet switches). For the simple case where only one Layer2 device is located between Grnet and the subscriber, we use scripting to query the speed and bandwidth settings at each L-2 border interface. We then reflect the speed setting of the border interface into a traffic shaping queue for the respective VLAN at the L-3 border. Using this technique, we make sure that the congestion points occur only at the L-3 border.

With the advent of hybrid networks and the tendency to carry high speed network traffic at the lowest layer possible (in order to avoid handling it with costly Layer 3 equipment), this part of current and future network is bound to expand. Whether this Layer 2 part of the network forms multiple paths between the connected Layer 3 devices (in which case the need for spanning tree algorithms arises in the common Ethernet case) determines in large part the complexity of the Layer 2 QoS solution that will have to be adopted.

Therefore, in designing and implementing the service described in this paper, we took into account the current need

for controlling traffic behaviour at the edge of the GRNET network (where it slips from current Layer 3 QoS model) and we also considered the increasing importance of that part of the network to the overall network architecture in the future.

III. Implementation Issues

IEEE 802.1Q (also known as VLAN tagging) defines a 3-bit field called Class of Service (CoS), which can be used in order to differentiate traffic. Table 1 shows the 8 possible values of the CoS field and their original purpose:

TABLE I. CoS field values

CoS	Acronym	Purpose
0	BE	Best effort
1	BK	Background
2	-	Spare
3	EE	Excellent Effort
4	CL	Controlled Load
5	VI	"Video" < 100 ms latency and jitter
6	VO	"Voice" < 10 ms latency and jitter
7	NC	Network control

For the purposes of our deployment, we have adopted the usage of CoS value 5 for marking premium traffic (which requires quality of service), CoS 0 for best-effort traffic and CoS 1 for less than best effort traffic. Traffic is marked as less than best effort when it is of minor importance, and is allowed to occupy at most 1% of the total bandwidth.

In the case of the GRNET [2] network, end to end traffic between client network interconnected through GRNET will traverse a combination of Layer 2 (switches) and Layer 3 devices (routers). To this end, the policies of the edge routers of the GRNET network must be adapted so that ethernet frames belonging to premium traffic are marked with CoS 5 at the output. Additionally, the port of the subscriber's switch which is connected on the edge router has to be configured in order to trust the values of CoS of the received traffic streams. Because CoS is part of the standard 802.1Q [5], the port on which the edge router is connected must be in trunk mode. When a port is in trunk mode it uses the tagged frames of 802.1Q [5] to communicate, which contain CoS and other information about virtual bridged local area networks (VLANs).

The procedure of deploying Layer 2 Quality of Service is quite similar to the one of Layer 3 QoS. Classification procedure is applied in incoming packets along with policing functions. Next, if traffic is in profile it is marked accordingly, else the packet is marked down or dropped. Next, the packets enter the switch's queues according to their markings.

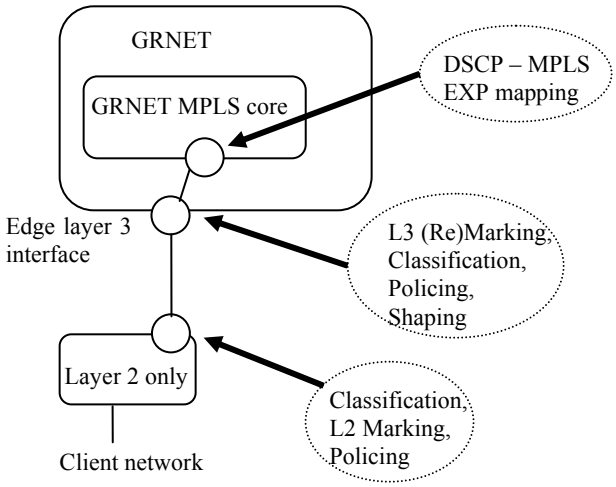

Figure 4. Schematic of L3 and L2 QoS actions

Queue management and scheduling are the most important issues in configuring Layer 2 Quality of Service. L2 Ethernet switches support a number of ingress and egress queues (switches in our testbed support 2 ingress queues and 4 egress queues). Scheduling in our equipment (Cisco Systems devices) is performed using the Shaped Round Robin (SRR) algorithm. The ingress queues can only be shared whereas the egress queues can also be shaped. When queues are shared their bandwidth is guaranteed to configured weights but is not limited to it. When a queue is empty, the other queues in shared mode share its unused bandwidth. When a queue is shaped it is guaranteed a percentage of bandwidth but it is rate limited to that amount. By default, from the ingress queues the second one is used to handle high priority traffic, and from the egress queues the first one is the high priority queue and it cannot be changed. Additionally, the high priority egress queue is by default shaped to occupy $^1/_{25}$ of total bandwidth, and when a queue is shaped any sharing settings are overridden. When the expedited output queue is enabled (as in our experiments, using the command **priority queue-out**), the expedited queue is serviced first until it is empty and then the other queues are serviced in a round-robin manner. More information can be found in [6]. In the GRNET network the edge routers shape the traffic on the output, so there is no need to shape the queues on the switches, however in our experiments, we use policies to limit the bandwidth when needed. Additionally, in the GRNET network the switch trusts the CoS of the packets coming from a GRNET edge router. By contrast, in our experiments traffic was classified by the switch and the DSCP field (46 for premium traffic, 0 for best-effort) was set, as in testing equipment policies that set CoS are not supported.

IV. Experiments

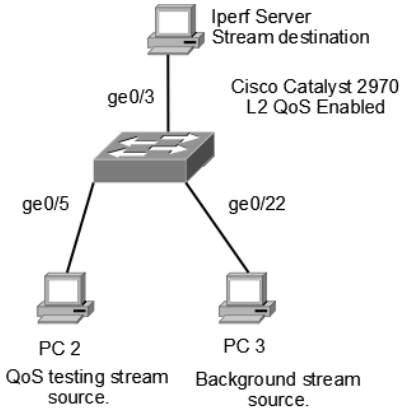

Figure 5. Topology of testbed

For our experiments, a gigabit switch Cisco Catalyst 2970 was used as well as 3 personal computers named PC 1, PC 2, PC 3 (Figure 5). PC 2 and PC 3 sent traffic to PC 1. PC 3 sent constant streams which represented the background traffic of a network and was serviced in a best effort manner. PC 2 sent premium traffic in a different manner in each test. Traffic was sent using the iperf tool [7], using UDP packets. Throughput, packet loss and jitter were measured. Although, the switch supports Gigabit Ethernet, our measurements were done using Fast Ethernet so that congestion conditions could be easily created. Iperf's statistics were produced at the server instance of the Iperf traffic generator and included the average throughput and the average jitter of the UDP traffic and the average throughput of the TCP traffic. Iperf calculates jitter using the RFC 3550 [13] definition that defines jitter as:

$$J_i = J_{i-1} + (\mid D(i-1,i) \mid - J_{i-1}) / 16$$

where D(i,j) is the difference of the interval between two successive packets at the receiver from the interval between two successive packets at the sender, defined as

$$D(i,j) = (R_j - R_i) - (S_j - S_i)$$

A. *Experiment 1: Lack of L2-QoS mechanism*

Our initial experiment was conducted before enabling any L2 QoS mechanism at the testbed switch, in order to have a reference point for our further evauations. In particular, PC 3 was sending a constant background stream of 100 Mb/s (representing aggregated low priority traffic that created congestion) and PC 2 was sending a foreground stream ranging from 1 to 8 Mb/s (representing traffic with low latency and jitter requirements). The sending algorithm of PC2 was to gradually increase the transmission rate from 1 to 5 Mb/s in 1 Mb/s steps and then peak at a transmission rate of 8 Mb/s. In this experiment it was clear that without preferential treatment, foreground traffic experienced the same packet loss and downgraded performance under congestion as the rest of the traffic.

B. *Experiment 2: L2-QoS deployment with exceeding packets dropped*

This experiment is the same as experiment 1 only that now QoS is applied at the foreground stream of PC 2. In this experiment the QoS stream was policed at 5 Mb/s and the policer was configured so that when the stream exceeded this rate, packets were dropped. In Figure 6 we can observe that there is no packet loss in the premium traffic stream until it reaches approximately 5 Mb/s (more specifically 4.85 Mb/s because of the protocol overheads). Then packet loss increases as excessive packets are dropped.

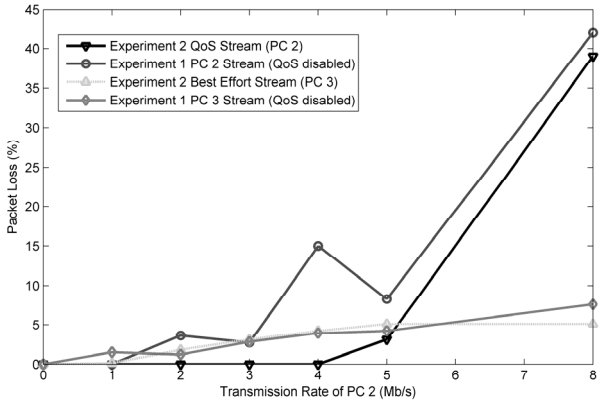

Figure 6. Experiments 1,2: Packet Loss – Transmission Rate

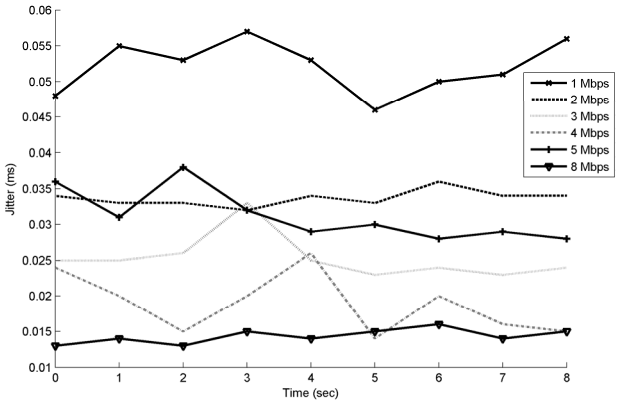

Figure 7. Experiment 2. Jitter – Time (QoS Stream)

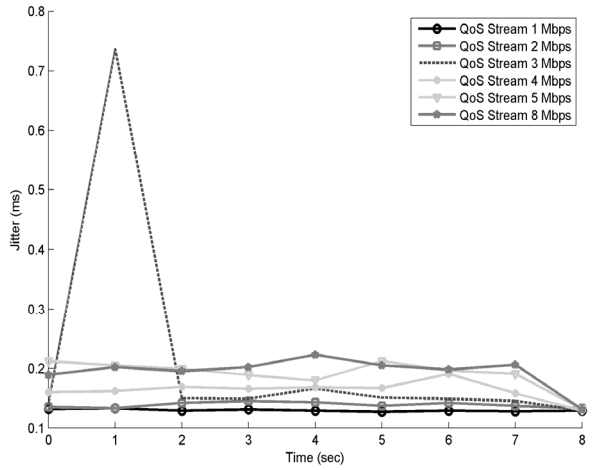

Figure 8. Experiment 2. Jitter – Time (Best Effort Stream)

Jitter is also much smaller for foreground traffic, and much more stable, which means that premium traffic packets arrive more orderly at their destination. Such a property is very desirable especially for real-time traffic (such as voice or video) which heavily depends on the timely delivery of successive packets.

C. *Experiment 3: L2-QoS deployment with exceeding packets remarked to Best-Effort*

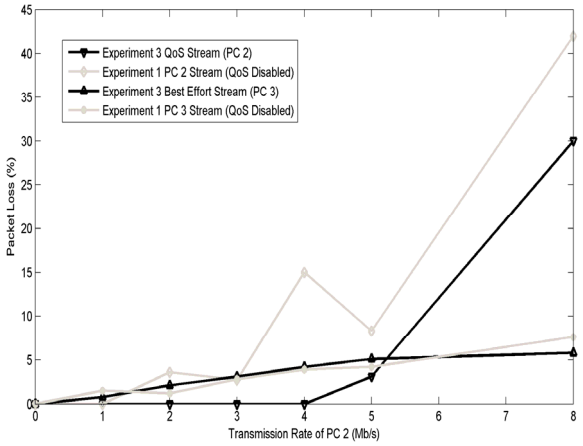

Figure 9. Experiments 1,3. Packet Loss – Transmission Rate

This experiment is the same as Experiment 2 with the only difference that in this experiment when the QoS stream exceeds 5 Mb/s the packets are not dropped but re-marked as best-effort. In Figure 9 it is evident that the QoS stream has no packet loss until it reaches approximately 5 Mb/s. Then the packet loss is increasing, following the pattern for best-effort traffic, which means that foreground traffic above 5 Mb/s is serviced through the same switch queues as background traffic.

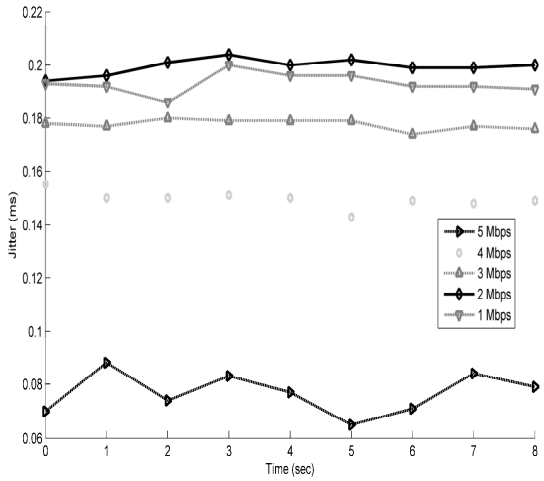

Figure 10. Experiment 3. Jitter – Time (QoS Stream)

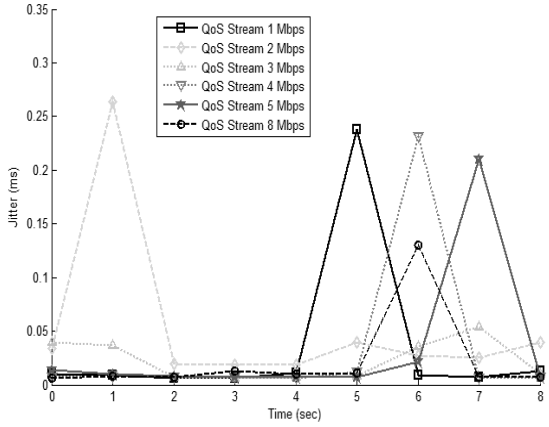

Figure 11. Experiment 3. Jitter – Time (Best Effort Stream)

As for jitter, our measurements show that remarked foreground traffic not only suffers from packet loss, but also from packet inter-arrival times which impact the measured jitter values.

In Figure 10 the 8 Mbps graph ranged from 2-17 ms and has been omitted to ensure better readability. These values are explained by the fact that the stream exceeds 5 Mbps and many of its packets are remarked. This causes some packets to arrive with more delay and out-of-order.

D. *Experiment 4: L2-QoS deployment with exceeding packets remarked to Best-Effort through Campus network*

This experiment is the same as Experiment 3 but now only PC 1 (Iperf Server) is directly connected to the switch, and traffic from PC 2 and PC 3 pass through the University Campus network as shown in Figure 12. The University network provides no quality of service whatsoever. Only the switch depicted provides quality of service to PC 2's stream.

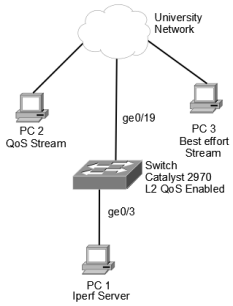

Figure 12. Topology of Experiment 4

As it can be seen on Figures 13,14 the jitter has now increased considerably in comparison with Figures 10,11 from the previous experiment. Additionally we can see that in average the best effort stream suffers from less jitter than the QoS stream but this is only because the experiments were run at different times, and traffic and congestion at the university network could not be controlled.

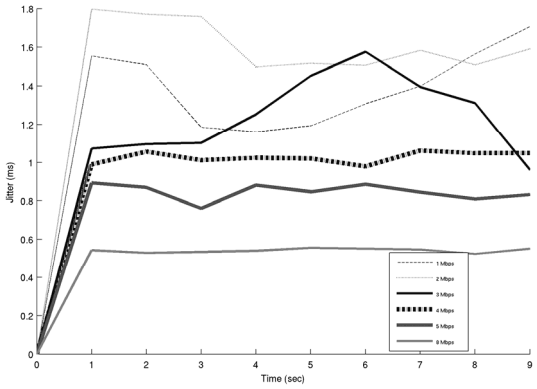

Figure 13. Experiment 4. Jitter – Time (QoS Stream)

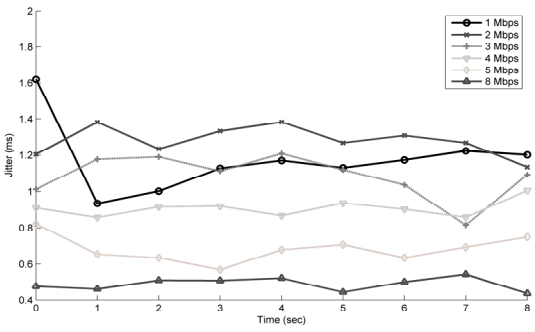

Figure 14. Experiment 4. Jitter-Time (Best Effort Stream)

V. Multiple L2 paths in Crete's MAN

An exception to the more common structure of the GRNET network described in section II is the part of the GRNET network at the island of Crete, which forms the Crete's MAN. It consists exclusively of L2 Ethernet switches which are aggregated to the only L3 device, a router at the city of Heraklio connected to the rest of GRNET (Figure 2). Some of the L2 interfaces are therefore considered part of the GRNET core network (the ones which form the MAN itself), while the rest connect to client networks, similarly to the common case discussed in previous sections. Therefore, for the latter case, the existing L2 approach can be still utilized. The core L2 devices form a ring consisting of 3 Ethernet switches (Cisco 3750), with several client networks connected on each one of them. Traffic between the client networks in Crete and towards the rest of the GRNET network is carried in VLANs in order to form isolated VPNs. A related limitation of the current Cisco L2 equipment is that it does not support QoS classification of traffic on VLAN ports, but only on physical ports.

Each VLAN has its own spanning tree which directs the traffic accordingly, and which can be quickly adjusted using Rapid Spanning Tree Protocol (RSTP) for link failure recovery and load balancing. In the case of a link failure, VLAN traffic using the failed link will be redirected due to the corresponding spanning tree protocol switching a blocking link's state to forwarding. This means that assuming the worst case scenario, a core L2 link will have to be able to carry the whole of the traffic traversing the core of Crete's L2 MAN. Under such an assumption, the worst-case dimensioning algorithm will have to allow premium traffic reservations up to the specified allocated percentage for the whole of the L2 MAN (conversely this can be expressed as the requirement that the allocated percentage should be calculated by adding all allowed traffic reservations through the MAN). The premium allocated percentage can follow the guidelines set by L3 allocations for L3 links of similar bandwidth. The symmetry of Crete' MAN regarding link capacity simplifies this calculation. The worst case assumption has also been the selected approach for premium reservations at the L3 part of the network, and is therefore a natural extension for this case.

VI. Conclusions and Future Work

In this paper we have presented a set of experiments and techniques that can be adopted so as to extend QoS from L-3 to L-2 at GRNET's network. We have described the service model and the different supported service types. Furthermore, we have discussed the issue of network congestion and how premium traffic streams can benefit from the application of QoS at L-2. We have described how the switch's queues can be set in order to provide QoS, as well as how the routers' configuration can be altered so as to provide marked flows of traffic.

As switching equipment becomes more and more powerful and versatile, we have moved the PHB and police-related functions from L-3 to the L-2 network boundary. Thus, by implementing a hybrid QoS scheme, using a translation from DiffServ to 802.1p, we can provide a unified QoS service layer across both the L-2 and the L-3 domains of GRNET. The conducted experiments acknowledged and proved that the activation of L2 QoS will benefit the overall result that is now produced by only L3 QoS in Grnet's network.

Our future plans include the application of L-2 QoS to the majority of GRNET's L2 equipment and in production's services portfolio. In addition we will enhance our QoS provisioning tool [14] with the necessary functionality and features in order to manage the L2 QoS service too.

REFERENCES

[1] Sven Ubik and Josef Vojtech "QoS in Layer 2 Networks with Cisco Catalyst 3550", CESNET Technical Report 3/2003.

[2] Greek Research Network (GRNET): www.grnet.gr

[3] C. Bouras, A. Karaliotas, M. Oikonomakos, M. Paraskevas, D. Primpas, and C. Sintoris, "QoS issues in the Research and Academic Networks: The case of GRNET", Industrial Conference on Multi-Provider QoS/SLA Internetworking (MPQSI 2005), Tahiti, French Polynesia, , 23 - 28 October 2005.

[4] A. Varvitsiotis, V. Siris, D. Primpas, G. Fotiadis, A. Liakopoulos, and C. Bouras, "Techniques for DiffServ-based QoS in Hierarchically Federated MAN Networks – the GRNET Case", The 14th IEEE Workshop on Local and Metropolitan Area Networks (LANMAN 2005), Chania. Island of Crete, Greece, , 18 - 21 September 2005.

[5] IEEE Standard for Local and Metropolitan area networks,Virtual Bridged Local Area Networks 802.1Q. http://standards.ieee.org/getieee802/download/802.1Q-2005.pdf

[6] Catalyst 2970 Switch Software Configuration Guide. Chapter 27: Understanding QoS. http://www.cisco.com/en/US/docs/switches/lan/catalyst2970/software/release/12.1_14_ea1/configuration/guide/2970SCG.pdf

[7] NLANR/DAST : Iperf - The TCP/UDP Bandwidth Measurement Tool. http://dast.nlanr.net/Projects/Iperf/

[8] F.K. Liotopoulos, and M. Guizani, "Implementing layer-2, connection-oriented QoS on a 3-stage Clos switch architecture", Global Telecommunications Conference, 2002, GLOBECOM '02. IEEE Volume 3, Issue , 17-21 Nov. 2002 Page(s): 2741 - 2746 vol.3

[9] Niclas Ek Department of Electrical Engineering, Helsinki University of Technology, "IEEE 802.1 P,Q - QoS on the MAC level", 1999, http://www.tml.tkk.fi/Opinnot/Tik-110.551/1999/ papers/08IEEE802.1QosInMAC/qos.html

[10] John Wernicke, "Simulative Analysis of QoS in Avionics Networks for Reliably Low Latency", Journal of Undergraduate Research,. Volume 7, Issue 2 - January/February 2006.

[11] Jacobs, A.; Wernicke, J.; Oral, S.; Gordon, B.; George, A., "Experimental characterization of QoS in commercial Ethernet switches for statistically bounded latency in aircraft networks", 29th Annual IEEE International Conference on Local Computer Networks, 2004,. 16-18 Nov. 2004 Page(s): 190 – 197

[12] Perez, J.A., Zarate, V.H., Cabrera, C., and Janecek, J., "A Network and Data Link Layer Infrastructure Design to Improve QoS for Real Time Collaborative Systems", International Conference on Internet and Web Applications and Services/Advanced International Conference on Telecommunications, 2006,. AICT-ICIW apos;06. 19-25 Feb. 2006 Page(s): 19 - 19

[13] RTP: A Transport Protocol for Real-Time Applications

[14] Grnet's Advanced Network Services Provisioning Tool http://anstool.grnet.gr