Evaluation of the Delivery QoS Characteristics of Gigabit Ethernet Switches

Răzvan Beuran[⋆‡], Mihai Ivanovici[⋆‡], Neil Davies[†], Bob Dobinson[⋆]

[⋆] CERN, European Organization for Nuclear Research, Geneva, Switzerland

[†] Predictable Network Solutions, Bristol, United Kingdom

[‡] University "POLITEHNICA", Bucharest, Romania

Abstract

The event selection system for ATLAS[1] is designed to perform real-time image processing on particle collision data equivalent to 2 TB/s. This data is filtered by a multi-level architecture, resulting in 200 GB/s of data analysed by a distributed system consisting of several thousand PCs and switches.

As part of our ongoing work on this system, we performed tests on several Gigabit Ethernet switches manufactured by market leaders, using our custom-built test equipment. We analysed the implications of running network devices at, and just beyond, saturation while deploying service differentiation mechanisms.

We quantified the quality degradation that traffic flows experienced when passing through switches. We focused on emergent properties in saturation, including fairness and fidelity to expectations. We discuss the ideals for switch behaviour and compare them against the observed behaviour of real implementations of differentiation mechanisms in switches. This creates a generic benchmark, which is independent of the switch internals.

Keywords: Gigabit Ethernet switch testing, QoS, emergent properties, saturation, scheduling algorithms, quality degradation.

1 Introduction

The Large Hadron Collider (LHC) is currently being built at CERN[2]. The proton bunches in LHC will cross at a frequency of 40 MHz [15], resulting in 10^9 events per second. The level-1 trigger will make the first level of event selection, reducing the initial event rate to at most 100 kHz. The high-level trigger must reduce the event rate further to about 100 Hz that are stored.

[1]One of the experiments being built at CERN[2].

[2]European Organization for Nuclear Research, Geneva, Switzerland.

Each event has a size of approximately 2 MB, leading to a total required processing capability of 200 GB/s and a storage capability of 200 MB/s.

The data is transported and analysed by means of a large network comprising thousands of PCs and switches. Control traffic and different phases of the data analysis will flow across the same infrastructure. It is expected that during normal operation saturation will occur, with the possibility of generating loss and additional delay. The data collection application can deal with a certain amount of loss, but in order to avoid interference with the underlying physics, loss of data is only acceptable as long as it is not biased. Continuous operation must be ensured, 24 hours per day for several years.

We believe it is compulsory to manage loss and delay at the network level, providing differentiated services to different applications (e.g. control, data collection, data analysis etc.). Given the unavoidable saturation and the complex nature of the network traffic, one of the prerequisites is to understand the emergent properties of switches. This also makes it possible to design a cost-effective solution that will bound the quality degradation in accordance with application requirements.

1.1 A novel view on Quality of Service

There are several well-known definitions for Quality of Service (QoS). ITU-T[3] considers it to be the "collective effect of service performances which determine the degree of satisfaction of a user of the service" [8]. However, the *de facto* meaning is: the performance characteristics of a network system (for which ITU-T uses the term "network performance").

We consider QoS to be *the fidelity of a system's observable behavior to expectations*. We measure and discuss the quality degradation in networks, that is the change in network service quality between two measuring points. We denote this degradation by the shorthand ΔQ. The total amount of degradation along a network connection is the aggregation of the local degradations that each sub-network and each network element (switches, routers etc.) on the way induces (see Figure 1). An essential property is that the experienced degradation only increases along a network path and cannot be undone. As a result, quality is only ever lost. A packet is either delayed or lost; a delayed packet cannot be made to arrive earlier, nor can a lost packet be recovered[4].

[3] International Telecommunication Union, Telecommunication division.

[4] Note that only data can be recovered by means of packet retransmission, not the lost packet itself.

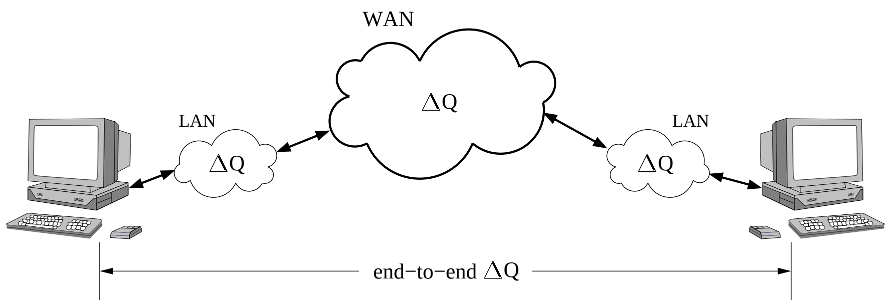


Figure 1: End-to-end quality degradation.

From a user perspective, understanding the application-level behaviour is essential. The end-to-end performance of a network application depends on the end-to-end network degradation. Each network application requires a minimum QoS level in order to run according to user expectations [2, 3, 4]. For example, for VoIP communication the total delay through a network must be less than 150 ms in order to have good interactivity [14], [18].

The contention for resources, typically due to saturation, is the prime cause of quality degradation in computer networks. By saturation we understand that particular situation when the demand for a resource approaches or exceeds its capacity. For a network running in saturation, such as the one deployed in ATLAS, the quality delivered to traffic flows deteriorates. Short term saturation causes an increase in delay, while long term saturation leads to loss. In order to handle such conditions, the emergent properties of the network must be predictable.

Our approach starts with the assessment of the emergent properties of switches, the building blocks of networks. A thorough study is required because even during low-average utilization, saturation conditions may occur. This is valid for ATLAS where the traffic resulting from bunch crossing events has a bursty nature [15] and also for the Internet traffic in general [10].

QoS mechanisms must be deployed in order to deliver differentiated quality to distinct applications. It is important to note that these mechanism are solely degradation sharing techniques and do not deliver quality by themselves. Since the total amount of degradation at a point in the network is conserved, QoS mechanisms may only be used to prevent degradation from exceeding certain bounds for the traffic of interest [14]. For example, the total amount of delay and packet loss can be shared by service differentiation between traffic flows in a controlled way.

1.2 A general benchmark for service differentiation

This article presents results and conclusions on the emergent quality properties of per-node QoS mechanisms as implemented in Gigabit Ethernet switches manufactured by market leaders. Our approach to testing switching devices is complementary to the test methodology proposed by IETF[5] since we emphasize the delivery QoS characteristics of switches. We focus on the study of the measurable outcomes of QoS mechanisms and we quantify the quality degradation (ΔQ) experienced by differentiated traffic.

A standard framework for QoS testing is lacking at the moment. The methodology for testing switching devices is currently specified in various RFCs[6] [5], [6], [11], [12]. These RFCs define the types of traffic to be used for benchmarking switches, the parameters to be measured, pre-defined states to be tested etc. However, RFCs do not define the framework for studying traffic differentiation and the consequences of deploying QoS mechanisms. This is regrettable since QoS characteristics are another important criteria for comparing switches, in addition to switching *per se*.

There are several companies engaged in benchmarking the QoS mechanism effectiveness in various areas, like Miercom [13] and The Tolly Group [16]. The main motivation of their tests is commercial, since these are intended to support marketing. The tests are based on RFCs and typically specified by manufacturers, not looking for the problems. Therefore, an incomplete evaluation is performed. In addition, switch testing consists mostly of stress testing the switching fabric (e.g. sending traffic at line-speed, in a fully-meshed configuration). However, differentiated service for traffic flows is not emphasized.

The goal of our QoS tests on switches is twofold: (i) to quantify the degradation introduced by different scheduling mechanisms; (ii) to observe the fairness regarding the treatment applied to different packet flows sharing the same priority queue.

Our work was partially motivated by the existing general belief that switches work as advertised. However, this is not always the case: "Vendors often engage in 'specsmanship' in an attempt to give their products a better position in the marketplace. This usually involves much 'smoke & mirrors' used to confuse the user" [5].

2 Switch Testing

General setups for benchmarking switches are described in [6]. Our setup comprises senders that generate artificial traffic with controlled properties (offered load, packet size etc.). The traffic flows through the device under

[5]Internet Engineering Task Force.
[6]Request For Comments from IETF.

test. A receiver collects this traffic and measures the QoS parameters in real time.

This setup is not only appropriate for testing the switching capabilities of network elements as described in IETF RFCs, but also for analysing the behaviour of switches when scheduling algorithms are deployed. The buffer management scheme was Tail Drop in all the tests we performed, i.e. when queues are full packets are dropped on arrival.

Other QoS features that we studied are policing and traffic shaping, complementary mechanisms for service differentiation that can be tested with our system.

2.1 Testbed

Our experiments were performed using the "Advanced Network Tester" developed in our laboratory [1]. This is a versatile test system that allows for highly accurate one-way delay measurements, with a precision of 200 ns. It can also measure in real time packet loss and throughput. This tester makes use of Alteon programmable PCI[7] Gigabit Ethernet network interface cards (NICs). Custom-designed PCI clock cards connected by short coaxial cables were used to achieve clock synchronization between senders and receiver, ensuring a global time reference for our measurements. This synchronization mechanism cannot be applied for remote locations. When testing long-haul connections, synchronization can be performed using GPS[8], as described in [9].

The setup we used for our experiments is shown in Figure 2. With this apparatus we measured the quality degradation experienced by different traffic flows when switch QoS mechanisms are deployed. Since the ATLAS experiment is not yet operational, we used traffic that drives switches into saturation, which is the key point where quality degradation occurs.

The traffic flows in our tests were generated by eight sources that send packets to the same destination. All packets from a source have the same priority, marked in the VLAN tag field of the Ethernet frame. Priorities are distinct between sources. The traffic we used to drive switches into saturation was Constant Load [5] (traffic with a constant inter-packet gap and a constant packet size) or Poisson (i.e. traffic with a negative exponential inter-packet gap distribution), characteristic for the ATLAS network scenarios.

[7] Peripheral Component Interconnect.
[8] Global Positioning System.

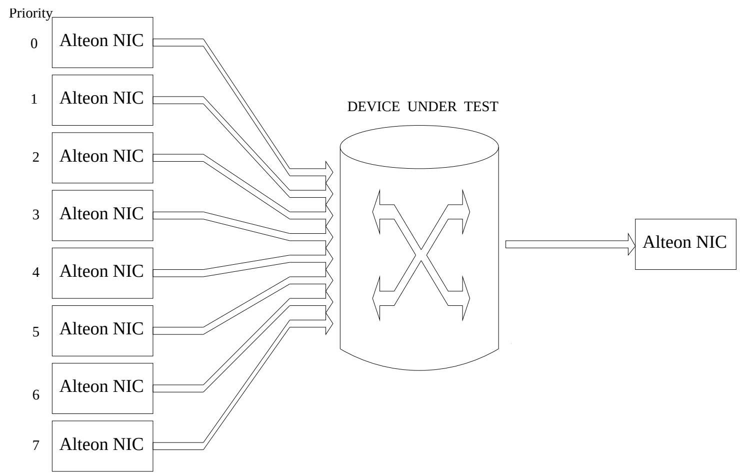


Figure 2: Test setup.

2.2 Expected behaviour of scheduling mechanisms

Strict Priority (SP) and Weighted Round Robin (WRR) are two scheduling mechanisms available in most Gigabit Ethernet switches. Each of them enforces service differentiation so that certain traffic classes are treated differently with respect to others. Their ideal, theoretical behaviour can be determined by considering a switch with no input buffers, an infinitely fast switching fabric and no contention for the access to the output buffers.

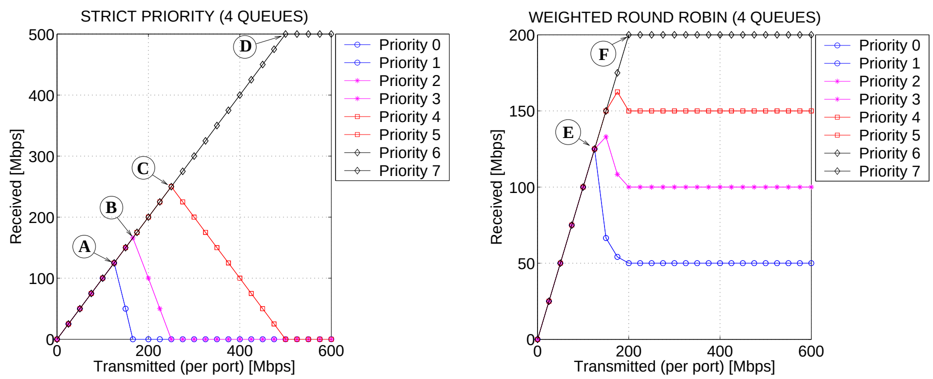
To convey differentiation between traffic flows we made use of the priority field in the VLAN tag of the Ethernet header, which allows for eight classes of traffic. The priority field is used by the switch to select the queue in which the packet will be placed. For switches with eight queues the mapping is one-to-one. If only four priority queues are available, then the VLAN tag priority is mapped to the service queue as follows:

Table 1: Priority mapping to service queues.

VLAN priority	0 & 1	2 & 3	4 & 5	6 & 7
Service queue	q_0	q_1	q_2	q_3

2.2.1 Strict Priority.

In SP [17] higher priority queues have precedence over the lower ones: a queue is not serviced as long as there are packets in higher priority queues. This may lead to starvation of the lower priority traffic flows when the offered load of higher priority traffic reaches or exceeds the available bandwidth. The induced packet loss and delay lead to a steep decrease in the



(a) (b)

Figure 3: Expected behaviour of SP scheduling (a) and WRR scheduling (b). For WRR, the weights are 0.4, 0.3, 0.2 and 0.1 for q_3 to q_0 respectively.

user-perceived quality for applications [2] and may even lead to application failure.

In Figure 3(a) we show the expected ideal behaviour for SP scheduling, in terms of throughput. We assume the case of a switch with Gigabit Ethernet ports and four service queues, the mapping being done according to Table 1. The lowest priority queue is q_0 and q_3 is the highest priority queue.

The points that define the behaviour of SP were denoted by (A), (B), (C) and (D). They indicate the place where the decrease in quality starts, as experienced by the traffic in queues q_0 to q_3, respectively. Point (A) represents the moment when the total offered load for all queues (q_0 to q_3) reaches 1 Gbps. Past (A), q_0 starts experiencing loss. Point (B) indicates that the total load for queues q_1 to q_3 is 1 Gbps; thereafter, q_1 loses packets as well. Point (C) shows the moment when the total amount of traffic in queues q_2 and q_3 reaches 1 Gbps. From this onward, packet loss occurs for q_2. Starting at point (D) q_3 is saturated and the received traffic is limited to 500 Mbps for each of the priorities 6 and 7. Everything else is lost.

2.2.2 Weighted Round Robin.

WRR is a variant of the Round Robin mechanism [7] that services each priority queue in a cyclic manner, proportionally to its associated weight. This allows control of the minimum amount of bandwidth guaranteed for each priority queue. The average delay for each traffic flow is bounded and can be estimated *a priori* (see Table 5).

In Figure 3(b) we show the expected ideal behaviour of WRR scheduling (a work-conserving mechanism). Again, we assume the case of a switch with Gigabit Ethernet ports and four priority queues. The corresponding weights are as follows: 0.4 for q_3, 0.3 for q_2, 0.2 for q_1 and 0.1 for q_0. Saturation is reached when all transmitters send traffic at 125 Mbps (point Ⓔ), thereafter service differentiation appears. Starting at point Ⓕ bandwidth guarantees are enforced.

3 Switch Evaluation

For each studied scheduling algorithm we compared its expected behaviour against that observed in our tests. Note that the performance of a switch depends significantly on the efficiency of the QoS mechanism implementations, its faithfulness with respect to the algorithm, the internal switch architecture etc. The benchmark we propose is however general and independent of these aspects, taking into account only the measurable outcomes of the tests.

We present here the results of our experiments on SP and WRR. For each traffic flow we performed basic ΔQ measurements: average throughput, average per-packet delay and average loss rate. All parameters were measured in steady state. Total offered load at the receiver was varied from 0 to 4.8 Gbps (each transmitting port sends up to 600 Mbps). The offered load was simultaneously modified for all transmitters. The results presented below were obtained with Constant Load traffic; they are very similar to those obtained when using Poisson traffic, at and beyond saturation of the output port.

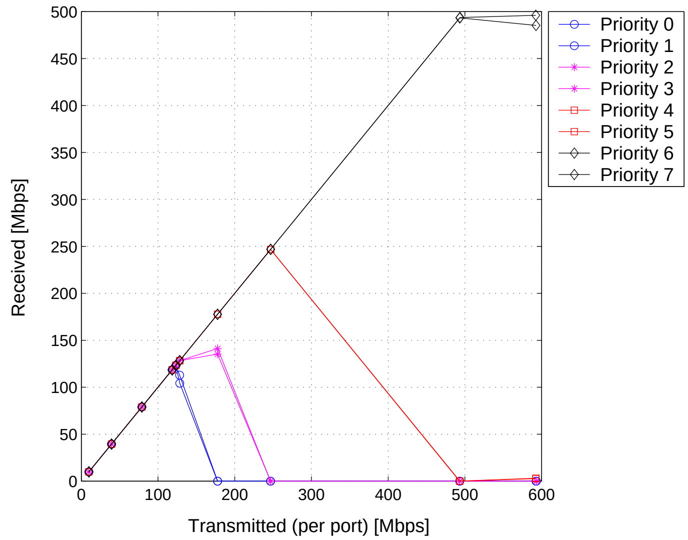
3.1 Strict Priority

In Figure 4(a) we show the results for a switch whose SP scheduling behaviour is very close to the ideal one. Note that lower priorities are starved as soon as higher priority traffic occupies the available bandwidth (i.e. 1 Gbps). Fairness is observed for the two traffic flows that share the same priority queues. However, for the highest priorities, at 600 Mbps, a slightly unfair treatment ($\approx 2\%$) is observed.

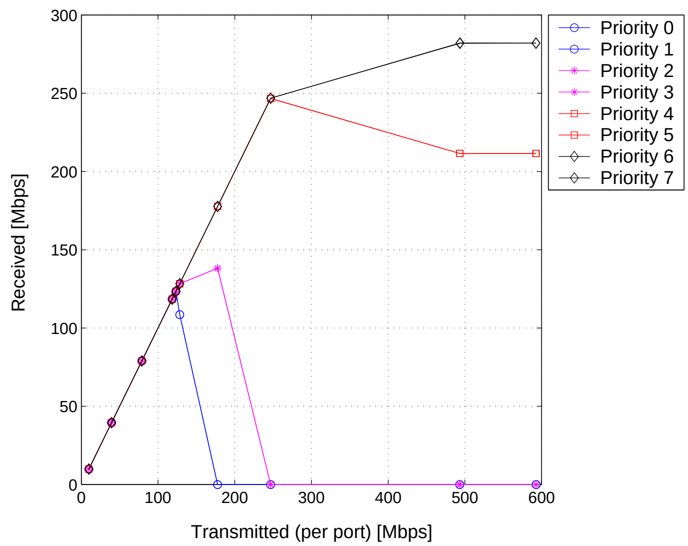
Figure 4(b) depicts the results of tests on a switch displaying a bad SP behaviour. The traffic having high priority is not starved for rates exceeding 500 Mbps; as a consequence the highest priority traffic doesn't occupy the entire bandwidth and loses packets (220 Mbps per sender at 500 Mbps offered load).

3.2 Weighted Round Robin

In Figure 5(a) the results for a switch with a good WRR behaviour are presented. Still, the sum of received traffic for all priorities is not 1 Gbps but only 990 Mbps. This means that 1% of the bandwidth is wasted, probably due to the increased complexity of the WRR algorithm compared to SP.



(a)



(b)

Figure 4: Example of good (a) and bad (b) SP behaviour.

The weights for the priority queues were: 0.4 for the highest priority, 0.3 for the high priority, 0.2 for the low priority and 0.1 for the lowest priority.

In the tests we performed, the WRR implementations work well when all traffic flows have same size packets. However, when using different sized packets, there is poor control of the bandwidth since scheduling is done on a per-packet basis. The solution is to schedule based on byte quanta, e.g. blocks of 256 bytes. In this way, a better control of the bandwidth may be achieved, its granularity being given by the size of quanta.

Nevertheless, we have established that bandwidth control problems still occur when testing switches that use 256-byte quanta. In Figure 5(b), we present the results of these tests. The packet sizes were 256 bytes, 512 bytes, 1024 bytes and 1518 bytes for the highest, high, low and lowest priority traffic flows, respectively.

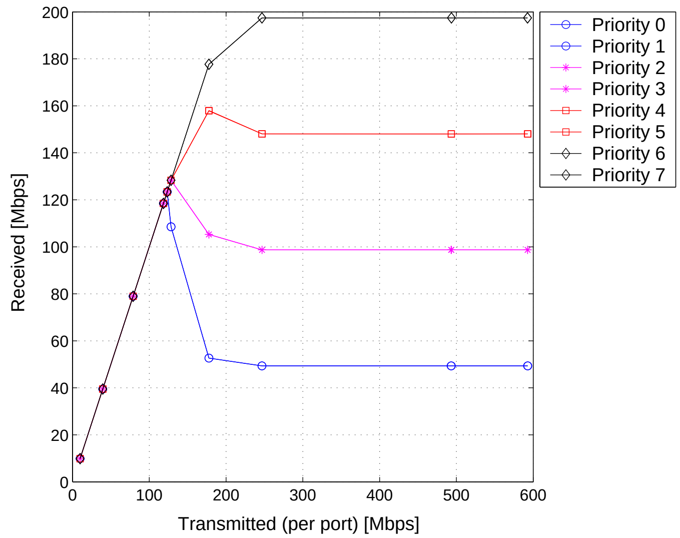
The behaviour in Figure 5(b) is obviously far from ideal, since the configured weights are not at all respected. The total bandwidth occupied by the received traffic is 963 Mbps, hence 37 Mbps are wasted; this may be due to the increased overhead of processing more small packets to achieve the same throughput.

3.3 Switch Comparison

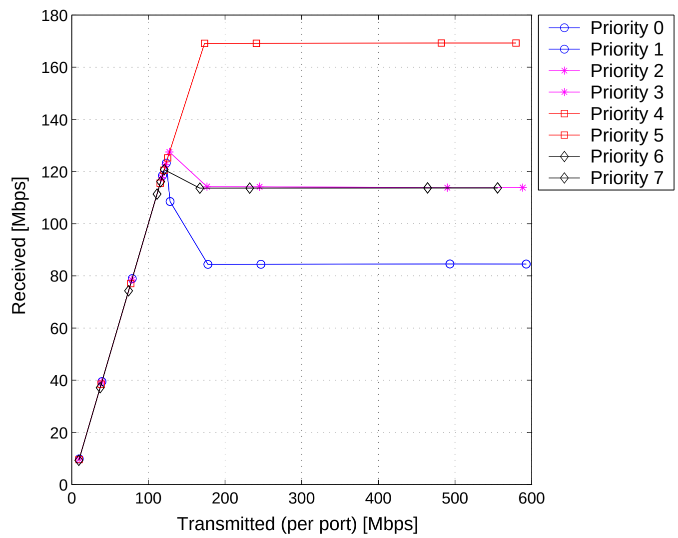
The graphs we presented capture a synthetic, but qualitative representation of the observed behaviour. A more precise, quantitative evaluation must be performed in order to be able to compare different switches.

We propose the computation of the achieved average delay and throughput ratios. Comparing these measures with the expected values derived from modeling and parameterized by the configured weights provides a valuable insight on how close the observed behaviour is to the expected one. This also makes it possible to choose the switch which provides (lower) worst-case bounds for the average delay.

Tables 2 and 3 show the average throughput (including the amount wasted) and average delay for SP scheduling. The values are obtained when each transmitter sends at 600 Mbps, therefore the switch is beyond saturation even for the highest priority service queue (q_3). The expected values for throughput and delay are included. We denote by $\Delta(\#n)$ the switch-dependent value of the average delay for q_3, where $n = \overline{1,4}$. In Table 3 we use ∞ to represent the delay of starved queues, for which packets never reach the destination.



(a)



(b)

Figure 5: Example of good (a) and bad (b) WRR behaviour.

Table 2: Throughput comparison for SP (600 Mbps per transmitter).

	Throughput [Mbps]				
Switch	q_0	q_1	q_2	q_3	Wasted
Expected	0	0	0	1000	0
#1	0	0	420	566	14
#2	0	0	422	564	14
#3	0.2	3.4	56	922	18.4
#4	0	0	5.4	972	22.6

Table 3: Delay comparison for SP (600 Mbps per transmitter).

	Delay [ms]			
Switch	q_0	q_1	q_2	q_3
Expected	∞	∞	∞	$\Delta(\#n)$
#1	∞	∞	4.9	3.5
#2	∞	∞	3.53	2.42
#3	3133	209	13.13	0.89
#4	∞	∞	351	2.43

From table 2 we observe that the first two switches fail to provide the highest priority with the best service (q_2 gets 42% of the bandwidth). Switch #4 has the best behaviour, even if q_2 is not completely starved, but 22.6 Mbps are wasted.

Tables 4 and 5 show the average throughput ratio, the wasted bandwidth and the average delay ratio for WRR scheduling. The scheduling is done on a per-packet basis and the size of all packets is 1518 bytes. The measured values were obtained when each transmitter sends at 600 Mbps. The expected values are computed based on the WRR weights (0.1, 0.2, 0.3 and 0.4 for switches #1 and #2; 0.1, 0.1, 0.3 and 0.5 for switches #3 and #4). We assumed that all service queues have the same length, a well-founded hypothesis based on our experience.

Table 4: Throughput comparison for per-packet WRR (1518 byte packets, 600 Mbps per transmitter).

	Throughput ratio				Wasted
Switch	q_0	q_1	q_2	q_3	[Mbps]
Expected	0.10	0.20	0.30	0.40	0
#1	0.10	0.20	0.30	0.40	14
#2	0.10	0.20	0.30	0.40	14
Expected	0.10	0.10	0.30	0.50	0
#3	0.08	0.08	0.33	0.50	14
#4	0.10	0.10	0.30	0.50	10

Table 5: Delay comparison for per-packet WRR (1518 byte packets, 600 Mbps per transmitter).

	Delay ratio			
Switch	q_0	q_1	q_2	q_3
Expected	0.52	0.27	0.15	0.06
#1	0.58	0.26	0.10	0.06
#2	0.60	0.27	0.09	0.05
Expected	0.53	0.32	0.11	0.04
#3	0.41	0.41	0.11	0.07
#4	0.28	0.28	0.20	0.23

Note in Table 4 the good agreement between the expected and observed values for the average throughput ratios. Switch #3 exhibits a deviation from the expected ratio of at most 0.03. All switches waste about 1.5% of the bandwidth.

The agreement concerning the expected and measured values of the average delay ratios is less good (see Table 5). Switches #1 and #2 have higher delay ratios (by 15%) for the lowest-weight service queue (q_0), meaning that it gets worse service than it should from the point of view of delay. On the other hand, switch #3 offers better service to q_0 than it should, and worse service to q_1. For switch #4, the average delay ratios are almost equal, even though the average throughput ratios are different. Albeit this might be explained by the switch internal architecture, it makes the point that the average delays for WRR are unpredictable in the sense they are implementation and switch dependent.

4 Conclusions

In this article we present a novel view on QoS, analysed through the quality degradation to which traffic flows are subjected in a network. We focus on delivery QoS characteristics of network devices, i.e. their emergent quality properties.

We show our results from tests performed on several Gigabit Ethernet switches. Our main interest was to study the overloaded behaviour of switches when QoS mechanisms are deployed. This means not only whether the system fails or not, but how its delivery QoS characteristics vary. We compared the observed behaviour against the ideal behaviour. We discovered that there is a significant gap between the expected and the observed behaviour.

Differentiated traffic sharing the same priority queue is not fairly treated in many of the switches we tested. Higher priority traffic may lose bandwidth in favour of lower priority traffic in some strict priority scheduling implementations. As expected, WRR scheduling works well for flows with same packet size. However, our tests show that it fails when flows have different packet size, even when using more sophisticated WRR quanta-based scheduling. We conclude that using equipment manufactured by market leaders doesn't necessarily guarantee the expected level of performance.

Hence, thorough tests are mandatory in order to validate the components used to build networks, especially in high-speed environments with strict QoS requirements. We provide a general methodology that allows this evaluation regardless of switch internals. This approach will be used in the design of the specialized networks of the ATLAS data collection system, to compare switches and select the best performing ones.

5 Acknowledgments

We thank Brian Martin, Cătălin Meiroşu and Jamie Lokier for their contributions to the development of the Advanced Network Tester. Peter Thompson and David Reeve had valuable suggestions for shaping the final version of this article.

References

[1] F. R. M. Barnes, R. Beuran, R. W. Dobinson, M. J. LeVine, B. Martin, J. Lokier, and C. Meirosu, “Testing Ethernet Networks for the ATLAS Data Collection System”, IEEE Trans. Nucl. Sci. 49, No. 2, April 2002, pp. 516-520.

[2] R. Beuran, M. Ivanovici, B. Dobinson, N. Davies, P. Thompson, “Network Quality of Service Measurement System for Application Require-

ments Evaluation", International Symposium on Performance Evaluation of Computer and Telecommunication Systems, July 20-24, 2003, Montreal, Canada, pp. 380-387.

[3] R. Beuran, M. Ivanovici, "User-Perceived Quality Assessment for VoIP Applications", technical report (delivered to U4EA Technologies), January 2004.

[4] R. Beuran, M. Ivanovici, V. Buzuloiu, "File Transfer Performance Evaluation", Scientific Bulletin of University "POLITEHNICA" Bucureşti (accepted for publication), September 2004.

[5] S. Bradner, "Benchmarking Terminology for Network Interconnection Devices", RFC 1242, July 1991.

[6] S. Bradner, J. McQuaid, "Benchmarking Methodology for Network Interconnect Devices", RFC 1944, March 1999.

[7] H. J. Chao, X. Guo, "Quality of Service in High-Speed Networks", John Wiley & Sons, Inc., 2002.

[8] ITU-T, "General Aspects of Quality of Service and Network Performance in Digital Networks, including ISDNs", ITU-T Recommendation I.350, March 1993.

[9] K. Korcyl, G. Sladowski, R. Beuran, R. W. Dobinson, C. Meirosu, M. Ivanovici, M. L. Maia, "Network performance measurements as part of feasibility studies on moving part of the ATLAS Event Filter to off-site institutes" , First European Across Grids Conference, Santiago de Compostela, Spain, February 2003

[10] W. E. Leland, M. S. Taqq, W. Willinger, D. V. Wilson, "On the Self-similar Nature of Ethernet Traffic", ACM SIGCOMM, San Francisco, California, U.S.A., 1993, pp. 183-193.

[11] R. Mandeville, "Benchmarking Terminology for LAN Switching Devices", RFC 2285, February 1998.

[12] R. Mandeville, J. Perser, "Benchmarking Methodology for LAN Switching Devices", RFC 2889, August 2000.

[13] Miercom, http://www.miercom.com/.

[14] D. Reeve, "A New Blueprint for Network QoS", Ph.D. thesis, University of Kent at Canterbury, August 2003.

[14] V. Servis, "Measuring Speech Quality over VoIP Networks", The Tolly Group, December 2001.

[15] P. Jenni, M. Nordberg, M. Nessi, K. Smith, "ATLAS High-Level Trigger, Data Acquisition and Controls", Technical Design Report, July 2003.

[16] The Tolly Group, http://www.tolly.com/.

[17] Z. Wang, "Internet QoS, Architectures and Mechanisms for Quality of Service", Morgan Kaufmann Publishers, 2001.

[18] ITU-T Recommendation Y.1541, "Network performance objectives for IP-based services", ITU-T, May 2002.