INTERNATIONAL JOURNAL OF COMMUNICATION SYSTEMS
Int. J. Commun. Syst. 2002; **00**:1–16 *Prepared using* dacauth.cls *[Version: 2000/05/31 v1.0]*

Evaluating the expedited forwarding of voice traffic in a differentiated services network

Artur Ziviani [1,2] †, José F. de Rezende [2], and Otto Carlos M. B. Duarte [2]

[1] *LIP6 – Université Pierre et Marie Curie,*
4, place Jussieu,
75252 - Paris Cedex 05 - France
Artur.Ziviani@lip6.fr
[2] *Grupo de Teleinformática e Automação (GTA)*
COPPE/EE - Universidade Federal do Rio de Janeiro
P.O. Box 68504 – 21945-970 – Rio de Janeiro – Brazil
{rezende, otto}@gta.ufrj.br

SUMMARY

The Differentiated Services architecture offers a scalable alternative to provide Quality of Service (QoS) to the new multimedia applications in the Internet. This paper aims at evaluating the delay and jitter experienced by voice traffic when handled by the Expedited Forwarding (EF) scheme. The analysis includes the effects of different packet scheduling mechanisms implementing EF and of the voice packet size. We also evaluate how efficiently each type of traffic uses an extra allocated bandwidth and the impact of traffic shaping. The results show that increasing the service rate share allocated to the EF aggregate does not significantly affect the competing best effort (BE) traffic. This holds as long as the BE traffic can use the bandwidth left unused by the EF traffic in idle periods.

KEY WORDS: Differentiated Services, Voice Traffic, Quality of Service

1. INTRODUCTION

An increasing number of private and commercial users rely on the Internet to communicate and exchange data. Nevertheless, the best effort service model of the current Internet architecture provides no guarantees to achieve the diverse Quality of Service (QoS) expectations of users and new multimedia applications [1, 2]. Such expectations imply that the network guarantees a certain amount of bandwidth, maximum packet delivery delay, or minimum variations in the inter-arrival time of consecutive packets (jitter). Therefore, a growing need to extend the

Contract/grant sponsor: FUJB, CNPq, and CAPES/COFECUB.
†Artur Ziviani has a scholarship from CAPES/Brazil.

Internet architecture with service differentiation mechanisms arises to allow some users and applications to obtain a better service than others at a higher cost [3, 4].

In order to address the issue of QoS provisioning over the Internet, different approaches are currently being developed, such as adaptive applications [5, 6], QoS-based routing [7, 8, 9], and the Differentiated Services (DiffServ or DS) architecture [10, 11]. The key difference between the previously proposed Integrated Services (IntServ) architecture [12, 13, 14] and DiffServ is that while IntServ proposes an end-to-end QoS on a per-flow basis, DiffServ provides long-term service differentiation among aggregates of traffic from different users. The DS architecture achieves scalability by reducing complexity within the network. Core routers maintain per aggregate state instead of one state per flow, pushing the complexity of handling individual flows to the network edges [15].

In this context, a contiguous set of nodes that operate with a common set of service provisioning policies is called a DS domain. At each DS domain, services are performed by traffic conditioning at the edges and simple forwarding mechanisms therein. Thus, individual flows needing a similar service are aggregated at the DS domain boundary and such aggregates receive differentiated services across the DS domain. At every DS-compliant node, a Per-Hop Behavior (PHB) defines how packets belonging to each aggregate are forwarded. Currently, two forwarding mechanisms are standardized by IETF: the Assured Forwarding (AF) [16] and the Expedited Forwarding (EF) [17, 18] PHBs. The EF PHB guarantees that the aggregate's departure rate from any DS-compliant node is at least equal to a configurable rate. Therefore, the EF PHB can be used to build a low loss, low latency, low jitter, assured bandwidth service through DS domains.

Although recent papers [19, 20, 21, 22, 23] analyze distinct aspects of differentiated services, the transmission of voice traffic over a DS domain remains not well explored. Addressing the issue of voice traffic evaluation in a DS environment becomes specially important as the DS architecture is expected to be one of the building blocks for Internet Telephony deployments [24, 25]. This paper focuses on how voice traffic may benefit from being treated by the EF PHB across a DS domain. The voice traffic sources are either a constant bit rate (CBR) source or an On-Off source, which are both often used to model voice traffic. The assessment covers the different algorithms that could implement the EF service. For the considered scheduling mechanisms, we identify the sources of delay and jitter. The analysis covers the impact of traffic shaping, the influence of the size of voice packets, and how efficiently each type of traffic uses an extra allocated bandwidth.

This paper is organized as follows. Section 2 reviews current proposals for Differentiated Services, pointing out the Expedited Forwarding scheme, and discusses the related work. Section 3 describes the models and the topology used in the simulations. Section 4 presents the simulation results. Finally, Section 5 contains some concluding remarks.

2. EXPEDITED FORWARDING AS A DIFFERENTIATED SERVICE

The identification of a flow aggregate within a DS domain is performed by checking a new field called Differentiated Services (DS) in the IP header [10]. This field holds a Differentiated Services Codepoint (DSCP) which is mapped into a Per Hop Behavior (PHB) at each DS capable router. The PHB defines a treatment to be offered to a packet until it is forwarded or discarded on each DS-compliant node. Ingress nodes of a DS domain can identify flows individually, so that packets belonging to a microflow could be classified and marked with the corresponding codepoint of a specific flow aggregate. Therefore, there is one state per flow at the DS domain boundary and just one state per aggregate within the DS domain. This strategy allows the DS architecture to scale by pushing the complex tasks to the edges of a DS domain, leaving therein only per-aggregate classification and forwarding mechanisms.

Due to the best effort model of the current Internet architecture, queues along the path are the main reasons for packet loss, latency, and jitter of a traffic. The Expedited Forwarding (EF) PHB [17] intends to reduce the time spent in queues by an EF aggregate while within a DS domain. To meet this goal, nodes that implement the EF PHB must assure that, at every transit node, the aggregate's maximum arrival rate is less than that aggregate's minimum departure rate. This guarantee must prevail independently of the intensity of any other traffic attempting to transit the node. The EF service is specified by the desired peak bit-rate for a specific flow (or aggregate of flows) [26]. The user contract with the network is not to exceed the peak rate. The network contract is that the contracted bandwidth will be available when traffic is sent. Ingress nodes must strictly condition the incoming EF traffic. Such a conditioning mechanism should assure the aggregate's arrival rate at the interior nodes to be compliant with the contracted rate. Packets that exceed the contracted profile are dropped or delayed until they are within the service profile.

An EF PHB implementation could be performed by different queuing scheduling mechanisms. A priority queue served in priority round-robin (PRR) gives the desired behavior. Thereby, the priority queue is served as long as it contains a packet to be forwarded. The other queue is only served when the priority queue is empty. In this case, the absence of traffic policing or traffic shaping at the DS domain boundary may allow the use of resources that are not contracted. This can harm other traffic aggregates or even cause starvation. Traffic shaping concerns the regulation of the transmission rate of a flow (or of an aggregate of flows). Packets exceeding the allowed rate profile are delayed until they are in accordance with the contracted rate profile.

An alternative to PRR scheduling is to use a weighted round-robin (WRR) scheduler [27]. In this scheduling police, the relative weight assigned to each queue determines the proportional share of the output bandwidth each queue is served with. A single queue is allocated to the EF traffic. The share of the output bandwidth assigned to the EF queue must be at least the EF contracted rate. Here, a misbehaviored user cannot jeopardize other aggregates because of the existing isolation between them. However, the EF PHB definition may be violated if a proper traffic shaping and reshaping is not performed at the DS nodes. Quantitative compliance testing of the offered EF service [28] is needed to allow vendors, operators, service providers, and users to exploit adequately the EF service.

A sequence of DS-compliant nodes that implement the EF PHB can offer a specific edge-to-edge service to an aggregate provided that traffic conditioning mechanisms are located at the DS domain boundary. An example of an intradomain service based on the EF PHB is called Virtual Wire [29], which intends to mimic, from the point of view of the edge nodes, the behavior of a hard-wired circuit of some fixed capacity. Such a service should be suitable for applications like Internet Telephony [30]. Some of the current proposals for Internet Telephony [24, 31] suggest the adoption of differentiated services mechanisms for the forwarding of voice flows. The most relevant QoS parameters for voice traffic include delay, jitter, and packet loss rate. These metrics reflect the continuity of a voice flow [32]. For interactive voice traffic, however, packet loss is not as critical as delay or jitter. A maximum end-to-end delay requirement must be satisfied in order to keep a voice conversation interactive. There should be no or little impact if one-way end-to-end delay is below 150 ms [33, 34]. A small jitter (i.e. 50 ms at most) is also suitable to reduce the buffer capacity needed by the receivers to compensate these variations in delay and keep the playback at the receiver smooth.

In [29], Jacobson *et al.* focus on the jitter analysis of voice traffic transported across a DS domain where all nodes have a priority queue implementation of the EF PHB. The work presented in [17] reports a comparison of the jitter observed by traffic transported by different possible implementations of the EF PHB definition. The results demonstrate that a priority queue provides the lowest jitter when compared with WRR. In [35], voice sources, either CBR or On-Off, are evaluated when transmitted over the original DS structure [26]. In [36], we evaluate the performance of voice sources when handled by the EF PHB, without taking into account the effects of traffic shaping and considering fixed size voice packets. The performance of voice flows transmitted through a DS domain implementing both EF and AF PHBs is analyzed in [37]. The effects of different scheduling schemes and of different packet sizes on voice delay are studied in [38]. Although the results obtained focus on delay, additional delays introduced by a possible shaping of the incoming voice traffic are not considered.

This paper aims at evaluating the delay and jitter performance of the EF PHB when transporting voice traffic modeled by CBR and On-Off sources. The analysis includes both the priority queue and the weighted round-robin implementations of the EF PHB. We also evaluate how efficiently each type of traffic uses an extra allocated bandwidth and the influence on performance of the voice packet size. Moreover, we study the effects of traffic shaping. The shaping of the EF aggregate at the ingress node allows the maintenance of a smaller reservation of bandwidth across the DS domain. Such action is done at the cost of an additional delay for the voice traffic. We analyze the trade-offs between this additional delay and the bandwidth allocated inside the DS domain for different possible implementations of the EF PHB and different models for voice traffic.

3. SIMULATION MODEL

We implement our simulation model in a modified ns-2 simulator [39, 40]. An extension to the basic ns-2 package was developed to support Differentiated Services. The implementation of different PHBs is achieved through a combination of a variety of packet classification, packet marking, packet scheduling, and traffic conditioning mechanisms. We use some of such mechanisms from our modified ns-2 to implement the EF PHB.

Figure 1 presents the network scenario used in our simulations. A number of voice flows are forwarded as an aggregate through a DS domain by the EF PHB. Along its path, the EF aggregate is shaped at the ingress node of the DS domain and reshaped at the interior nodes in order to be at any moment in accordance with the contracted rate and not violate the EF definition. These shapers are implemented by a token bucket algorithm. The token bucket depth is always set to be equivalent to the size of one voice packet, that is, no burst exceeding the contracted rate is allowed into the DS domain. The token rate may have values that range from the average rate of the EF aggregate to the peak rate of this aggregate depending on the simulation.

The EF voice sources generate voice traffic following either the On-Off model or the Constant Bit Rate (CBR) model. When an On-Off source is in the On state, it generates voice traffic at

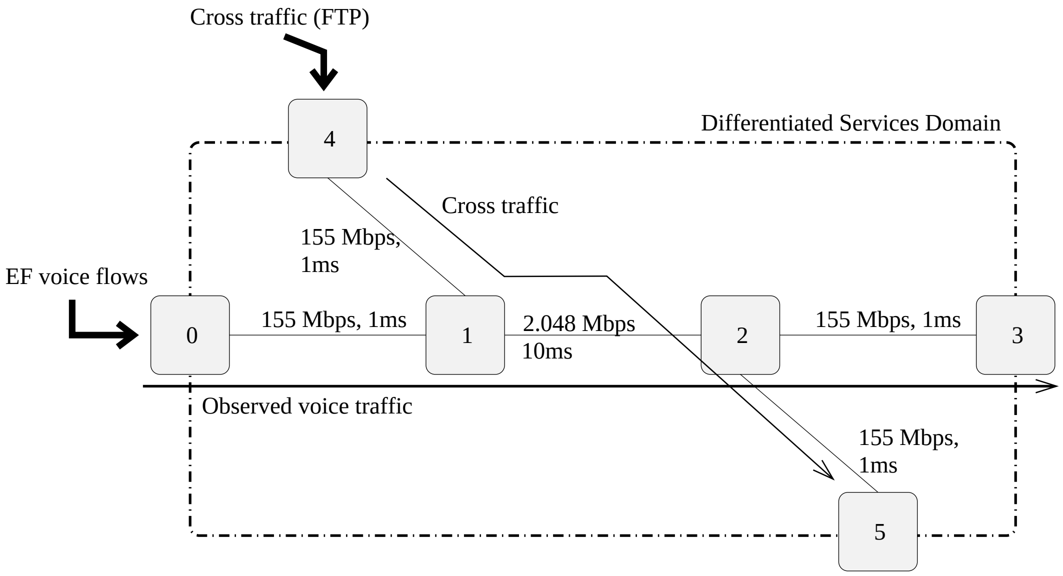


Figure 1. Network scenario.

a peak rate of 64 kbps. The On periods are determined by a random variable with exponential distribution with mean 400 ms. The Off periods are also exponentially distributed by a random variable with mean 600 ms. Therefore, the average transmission rate for each On-Off source is 25.6 kbps. Since the allocated share of the bottleneck link to EF traffic is 512 kbps, there are 20 On-Off sources, leading to a worst-case situation as their average rates exactly fill up the EF allocation.

The CBR sources have rates of 64 kbps with a small variation around this value. Such sources are called jittered CBR sources [35, 17]. This kind of sources represents better a real implementation of a digitized voice generated by an application running on a desktop computer. We use 8 CBR sources to fill up the 512 kbps share of the bottleneck link allocated to EF traffic. The voice packets may have different sizes ranging from 40 to 160 octets. Besides the EF voice flows, the bottleneck link receives cross traffic composed of FTP connections that have their packets forwarded as best effort packets. The TCP cross traffic sources use the TCP-Reno implementation which adds the fast recovery mechanism to the original congestion control mechanisms of the TCP. The cross traffic packets have 1500 octets which leads to a worst-case scenario for delay and jitter. Over 80% of Internet traffic packet sizes are much smaller than 1500 octets [41, 29]. Thus, a realistic traffic mix would give a better performance than the ones observed in our simulations. The TCP cross traffic is generated by FTP applications, active during the whole simulation. As the goal of this paper is to evaluate the delay and jitter performance of the voice traffic, the buffers of the shapers and of the DS-compliant nodes are large enough so that EF packets are never discarded due to lack of buffering resources.

4. RESULTS

The results are measured from one EF reference voice flow. That is, on all simulations the delay and jitter are from packets belonging to the reference flow within the DS domain. The delay of a single packet is the interval between the arrival of the packet at the ingress node and its departure from the egress node, including any delays introduced by traffic shaping. As in [17], we assume the definition of jitter as the absolute value of the difference between the arrival times (a) of two adjacent packets minus their departure times (d), then

$$jitter = |(a_j - a_i) - (d_j - d_i)|. \tag{1}$$

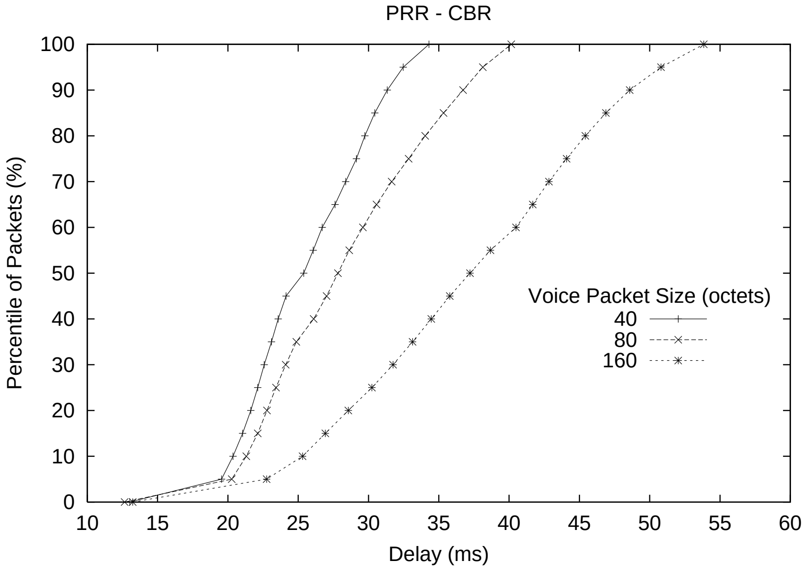
4.1. PRR Implementation of the EF PHB

First we study the service offered by a PRR implementation of the EF PHB to voice traffic modeled by the jittered CBR sources (10% of variation around an average rate). The voice traffic is shaped at the ingress node of the DS domain to the rate of 64 kbps. Figure 2(a) presents the percentile delay for three different voice packet sizes: 40, 80, and 160 octets. Figure 2(b) shows the corresponding results for jitter. The smaller the voice packet, the better the delay and jitter performance of voice flows. Moreover, a smaller voice packet also implies a smaller delay for the packetization and the depacketization of the voice samples at either end. However, there is a trade-off between the voice packet size and the overhead of the header.

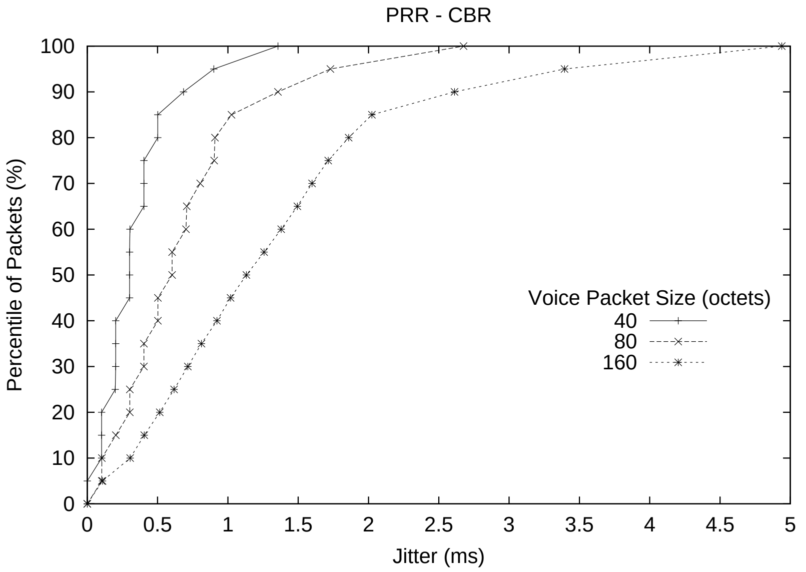
When one EF voice packet is enqueued at the shaper, it must wait the forwarding of all EF voice packets ahead of it in the queue before it is allowed to proceed. Depending on the size of EF voice packets ahead of one packet in the shaper's queue, it takes a larger or a smaller time for this packet to be forwarded. Also, if there is a low priority packet being transmitted when an EF packet arrives, it must wait until the low priority packet is completely served since the scheduling mechanism is nonpreemptive. It should be noted that the low priority packets are the TCP cross traffic packets having 1500 octets, thus leading to a worst-case scenario for delay and jitter. The difference in the individual delay of packets also influences on the jitter of the CBR traffic, as shown in Figure 2(b).

Next we evaluate the performance of a PRR implementation of the EF PHB when handling voice traffic generated by On-Off sources. The voice packets have 40 octets. As in the previous simulation, the EF aggregate is shaped at the DS domain boundary. We define the *individual flow shaping rate* as the resulting rate that would be obtained by each component of a flow aggregation if they were equally affected by the shaping of the aggregation. Therefore, the actual token rate of the EF aggregation shaper is the product of the individual flow shaping rate and the number of EF active flows. Figure 3 presents the percentile delay as a function of the individual flow shaping rate, which ranges between 28 kbps and 48 kbps for each flow. Similarly, Table I provides the average delay for the same range of individual flow shaping rates. In these simulations, the EF aggregate is composed of 20 individual flows for a worst-case situation.

The greater the individual flow shaping rate, the smaller the time spent by packets waiting to be in accordance with the contracted rate. That is, as the individual flow shaping rate increase, more bursts of the EF aggregate are allowed into the DS domain. The DS-compliant interior nodes forward these voice bursts with priority over the best effort traffic. Nevertheless,



(a) Delay (ms)



(b) Jitter (ms)

Figure 2. CBR traffic handled by a PRR implementation of the EF PHB.

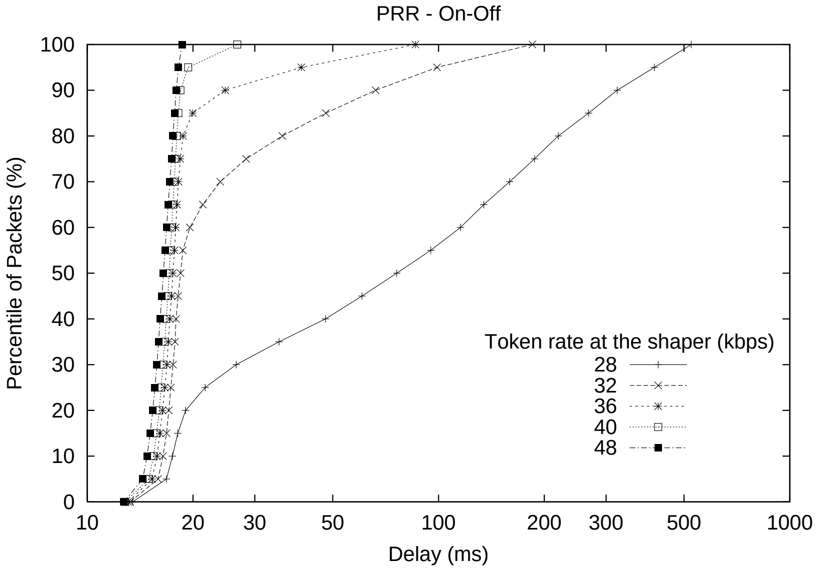


Figure 3. On-Off traffic handled by a PRR implementation of the EF PHB.

Table I. On-Off traffic delay performance in ms for a PRR implementation of the EF PHB.

Individual flow shaping rate at the shaper (kbps)	Delay (ms)
28	123.82
32	31.68
36	20.53
40	17.25
48	16.42

the total goodput of the best effort traffic composed of 8 FTP connections are not significantly affected by the variation of the allowed voice bursts, as shown is Figure 4. This is because the best effort traffic uses the idle periods of the bursty traffic generated by the On-Off sources. For an EF aggregate composed of 20 voice flows, the total goodput of the best effort traffic levels off slightly above 1.5 Mbps, which is reasonable since the share of the bottleneck link of 2.048 Mbps allocated to the EF aggregate is 512 kbps. For a similar worst-case situation, the allocation of a larger share of the bottleneck link to the EF aggregate decreases the value on which the total goodput of the best effort traffic levels off. We observe such result for EF aggregates composed of 40 and 60 voice flows, which correspond to approximately 50% and 75% of the bottleneck link, respectively. As the capacity of the bottleneck link allocated to the EF aggregate increases, the best effort traffic tends to starve.

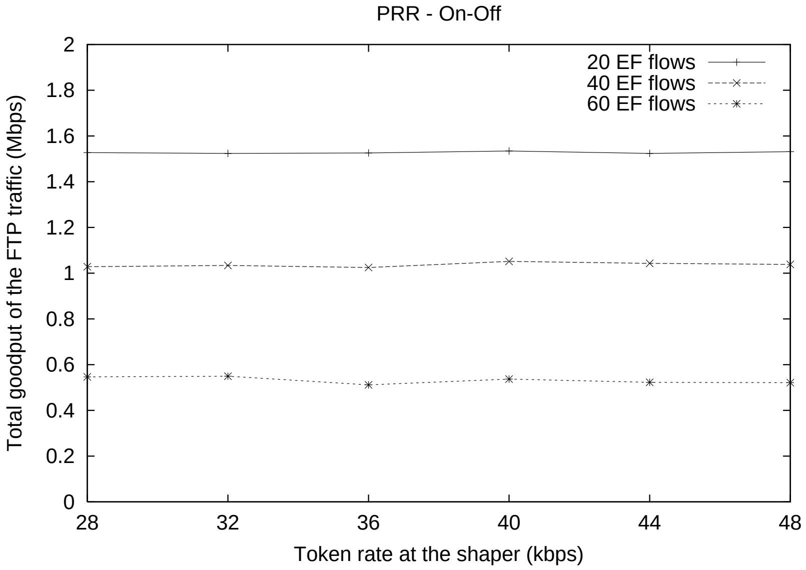


Figure 4. Goodput of the FTP cross traffic in PRR implementation of the EF PHB.

4.2. WRR Implementation of the EF PHB

A WRR implementation of the EF PHB is achieved by allocating to the EF aggregate a portion of the output bandwidth that is larger than the contracted rate to the EF aggregate at any moment in the DS domain. If the WRR weight is chosen to exactly balance arrival and departure rates the results are not stable [17]. A service-to-arrival rate ratio r is defined as

$$r = \frac{WRR_{share} \cdot BW_{output}}{PR_{EF}} = \frac{reserved\ bandwidth}{requested\ bandwidth}, \quad (2)$$

where WRR_{share} is the fraction of the output link bandwidth (BW_{output}) allocated to the EF aggregate, and PR_{EF} is the peak rate of the EF aggregate entering the DS domain.

When one has a service-to-arrival ratio greater than 1 in all links within a DS domain, one reserves more bandwidth than it was requested. The minimum service-to-arrival ratio to have stability is around 1.03 [17], and the value used in our simulations is initially set to 1.06, i.e., the actually reserved bandwidth is 6% larger than it has been requested.

In contrast to the PRR, where packet delays are mainly due to traffic shaping, the WRR implementation may have an additional source of delay and jitter. At any DS node, even if an EF packet is in accordance with the contracted rate, it may have to wait for its queue to be served by the WRR scheduler. If an EF packet arrives just after the EF's turn in the scheduler has passed, the EF packet has to wait the whole round allocated to the best effort traffic to be forwarded. Any EF packets arriving while the best effort traffic is being served are delayed. Furthermore, when the EF traffic is served the accumulated EF packets are forwarded in a

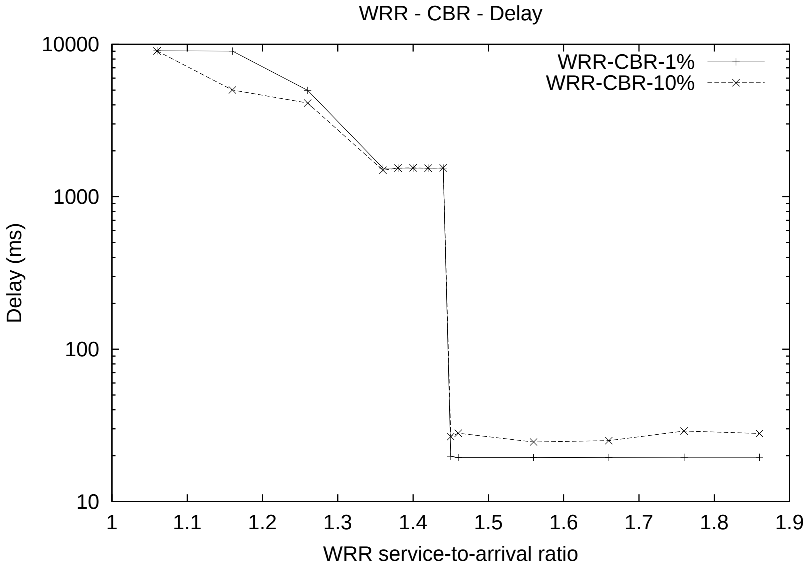


Figure 5. Average delay of CBR traffic handled by a WRR implementation of the EF PHB.

burst, which has to be conditioned by the downstream intermediate shapers. Such reshaping also introduces additional delays to the voice traffic, resulting in a cascade delay effect.

With the variation of r, it is possible to allocate a greater WRR service share to the EF aggregate. This reduces the average time spent by EF packets waiting for being scheduled at each DS-compliant node. That is, by increasing the WRR service-to-arrival ratio r, an extra bandwidth is allocated to the EF aggregate. Increasing the service-to-arrival ratio r means that the traffic must always be shaped upstream, otherwise misbehaving EF traffic could take more than its fair share.

The average delay of the CBR traffic composed of 40 octets packets and served by a WRR implementation of the EF PHB for different values for the WRR service-to-arrival ratio r is shown in Figure 5. The curves show results for jittered CBR sources with $\pm 1\%$ and $\pm 10\%$ variation. Both aggregates are shaped to the rate of $N \times 64$ *kbps*, where N is the number of flows that compose the aggregate. For both values of variation, the average delay is too high to interactive voice traffic when $r < 1.45$ because of the cascade delay effect with delays being introduced by the waiting in the scheduler and the traffic reshaping. When $r \geq 1.45$, the delay levels off around 20 ms for the 1% variation jittered CBR sources and around 27 ms for the 10% variation jittered CBR sources. Under the considered conditions, a 45% over-provisioned share in the WRR scheduler is needed to mitigate the delay imposed on voice traffic by the non-voice traffic, thus avoiding the cascade delay effect. Nevertheless, for instance, the goodput of the cross traffic only decreases from 1.645 Mbps for $r = 1.06$ to 1.528 Mbps for $r = 1.86$. This is possible because the best effort traffic uses the bandwidth of the EF aggregate when there are idle periods due to the work-conserving nature of the WRR scheduler.

Table II. On-Off traffic delay performance in ms for a WRR implementation of the EF PHB.

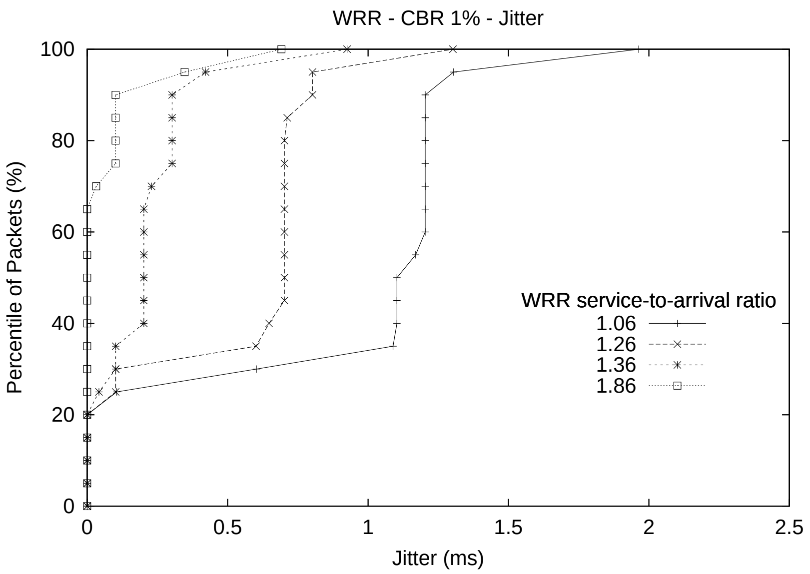
Individual flow shaping rate at the shaper (kbps)	WRR service-to-arrival rate 1.06	1.46	1.86
28	4,989.37	184.38	105.52
32	1,930.45	39.90	31.32
36	453.93	26.19	20.23
40	121.87	17.95	17.27
44	46.54	17.05	16.76
48	31.29	16.43	16.44

The jitter performance, however, continually improves as the WRR service-to-arrival ratio r increases. Figure 6 shows the jitter of the CBR sources while an extra share of bandwidth is allocated to the EF aggregate. For a variation of 1% around the average rate and for all applied values of r, most voice packets present a narrow range of jitter variation, causing thus the cliffs presented in Figure 6(a). Meanwhile, for a variation of 10%, the jitter of voice packets presents a worse performance and has a wider range of variation. How much the traffic are jittered at the source directly influences on the jitter variation. Such effect may be observed comparing the range of jitter variation in Figures 6(a) and 6(b).

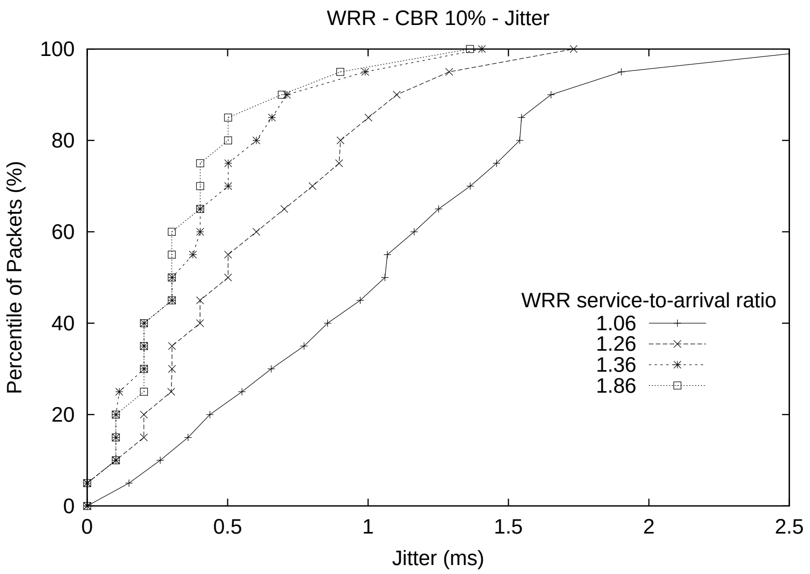
The delay and jitter performance for the voice traffic generated by On-Off sources and handled by a WRR implementation of the EF PHB is analyzed next. In this case, voice packets are delayed by two main reasons: shaping delay because of the bursty nature of On-Off sources and the waiting time in the queue served by the WRR scheduler. Hence, the delay performance may be improved by increasing the individual flow shaping rate at the shaper and/or the WRR service-to-arrival ratio.

Table II provides the results for the On-Off traffic composed of 40 octets packets as we increase both the individual flow shaping rate of the shaper and the WRR service-to-arrival ratio r. From Table II, the configuration with 28 kbps for the individual flow shaping rate at the shapers and $r = 1.06$ presents a prohibitive delay for interactive voice traffic. This configuration allows a goodput of 1.645 Mbps for the FTP cross traffic. At the other extreme of Table II, where we have 48 kbps as the individual flow shaping rate of shapers and $r = 1.86$, the delay becomes low enough to allow the feasibility of interactive voice traffic. It should be noted that just for a service-to-arrival rate ratio $r = 1.86$, the WRR gets a delay performance similar to PRR (see also Table I). Although the allocated bandwidth for the EF aggregate inside the DS domain is almost 1.8 Mbps ($48\ kbps \times 20\ EF flows \times 1.86$), which is nearly 87% of the bottleneck link, the goodput of the FTP cross traffic remains equal to 1.528 Mbps. This is possible because the best effort traffic uses the bandwidth of the EF aggregate when there are idle periods and the EF aggregate is limited to a 512 kbps share of the bottleneck link. Note that the average delay decreases from 4,989.37 ms to 16.44 ms at the cost of just 100 kbps from the goodput of the best effort traffic.

The jitter performance of the On-Off traffic handled by the WRR implementation of the EF PHB is shown in Figure 7. With a larger individual flow shaping rate at the shaper, there is a smaller influence on jitter of the WRR service-to-arrival ratio. This is illustrated when we compare Figures 7(a) and 7(b).

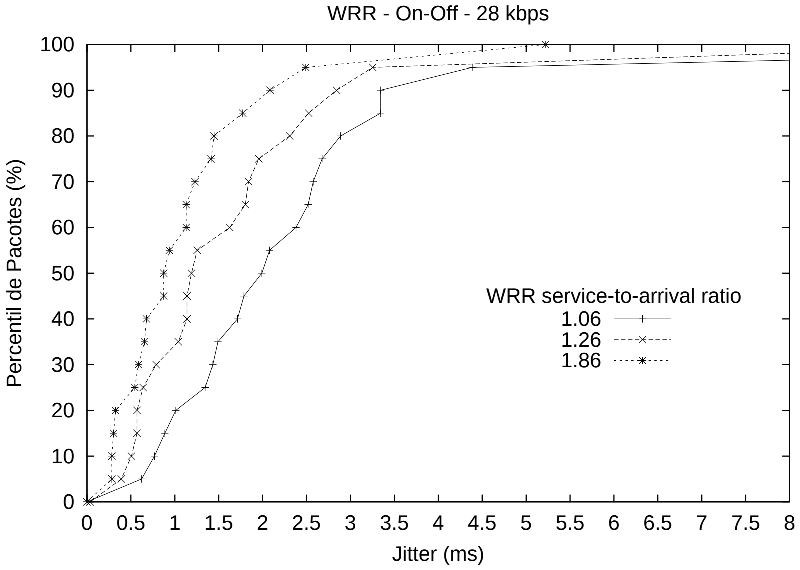


(a) 1% variation

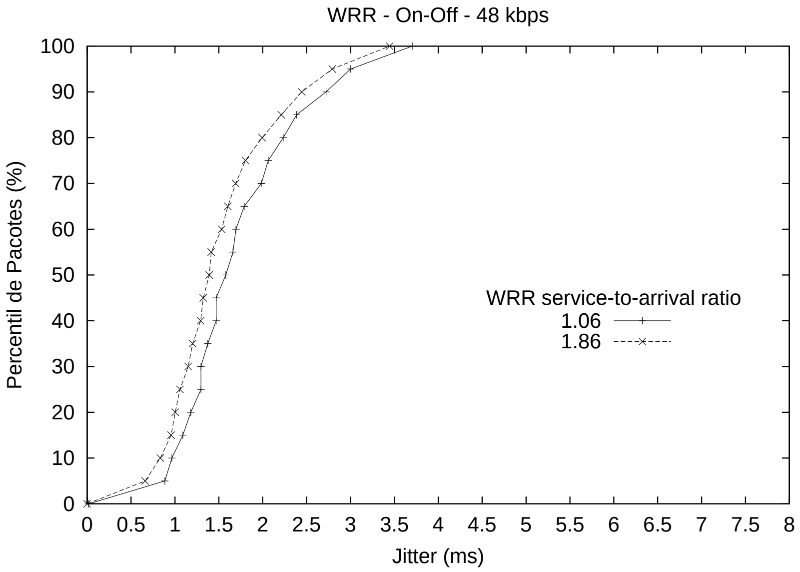


(b) 10% variation

Figure 6. Jitter of CBR traffic handled by a WRR implementation of the EF PHB.



(a) Individual flow shaping rate = 28 kbps



(b) Individual flow shaping rate = 48 kbps

Figure 7. Jitter performance of the On-Off traffic for a WRR implementation of the EF PHB.

5. CONCLUSION

In this paper, we present a simulation study of voice traffic over a Differentiated Services architecture for the Internet. We evaluate the service offered to voice traffic by the Expedited Forwarding PHB. The voice traffic was modeled by CBR and On-Off sources. We compare the performance obtained considering two different implementations of the EF PHB. One is a priority round-robin (PRR) and the other a weighted round-robin (WRR) scheduling mechanism. The analysis focuses on the delay and jitter observed by voice flows across a DS domain.

The traffic shaping of an EF aggregate composed of individual voice flows is fundamental to keep the delay and jitter within an acceptable bound. It also provides control over the incoming rate of the EF aggregate in order to guarantee that the EF definition is valid at any DS-compliant node. Although the traffic shaping introduces an additional delay, it allows the limiting of the contracted rate for the EF aggregate within a DS domain. The EF PHB is expected to constitute an expensive service. Thus, a trade-off may be established between allowing an additional but tolerable shaping delay and contracting a less costly amount of resources. The results show that as one increases the individual flow shaping rate at the shaper while keeping the same amount of EF flows, voice flows experience a smaller delay and jitter. Such an increase reduces the service rate share allocated to the best effort aggregate. However, the results also show that this does not significantly influence the best effort traffic performance. This behavior is due to the work-conserving nature of the schedulers, which allows the BE traffic to use the bandwidth left unused by the EF traffic in idle periods.

The PRR implementation of the EF PHB provides the best results for delay as well as for jitter. Nevertheless, the study of alternatives is needed since this scheduler has some undesirable features. The PRR cannot avoid low priority traffic starvation and does not offer a good isolation among different aggregates. Thus, improving the performance of a WRR implementation of the EF PHB is justified when is known that this scheduling mechanism is already available on commercial routers. Besides the shaping delay, the WRR implementation may introduce an additional delay to the EF voice flows due to the waiting it may impose on each queue to be served. To improve the performance of voice flows, one may increase the individual flow shaping rate to decrease the shaping delay. Similarly, one may increase the service-to-arrival rate ratio to decrease the retention delay at the WRR scheduler. These two parameters may be viewed as tuning knobs for adjusting the configuration of a WRR implementation of the EF PHB within a DS domain.

REFERENCES

1. Lefelhocz C, Lyles B, Shenker S, Zhang L. Congestion control for best-effort service: Why we need a new paradigm. *IEEE Network* 1996; **10**(1):10–19.
2. Xiao X, Ni L. Internet QoS: the big picture. *IEEE Network* 1999; **13**(2):1–13.
3. Clark DD. Adding service discrimination to the Internet. *Tech. rep.*, MIT, Sept. 1995.
4. Shenker S. Fundamental design issues for the future Internet. *IEEE Journal on Selected Areas in Communications* 1995; **13**(7):1176–1188.
5. Diot C, Huitema C, Turletti T. Multimedia applications should be adaptive. In *Proceedings of High Performance Computing Systems and Applications - HPCS'95*, Aug. 1995.
6. Amorim MD, Duarte OCMB, Pujolle G. Application-aware multicast. In *Proceedings of the IEEE Global Telecommunications Conference - GLOBECOM'01*, Nov. 2001.

7. Crawley E, Nair R, Rajagopalan B, Sandick H. A framework for QoS-based routing in the Internet. *Internet RFC 2386*, Aug. 1998.
8. Costa LHMK, Fdida S, Duarte OCMB. Developing scalable protocols for three-metric QoS routing. *Computer Networks* 2002; **39**(6):713–727.
9. Costa LHMK, Fdida S, Duarte OCMB. Hop by hop multicast routing protocol. In *Proceedings of ACM SIGCOMM'2001*, Aug. 2001.
10. Nichols K, Blake S, Baker F, Black DL. Definition of the differentiated services field (DS field) in the IPv4 and IPv6 headers. *Internet RFC 2474*, Dec. 1998.
11. Blake S, Black DL, Carlson M, Davies E, Wang Z, Weiss W. An architecture for differentiated services. *Internet RFC 2475*, Dec. 1998.
12. Braden R, Clark DD, Shenker S. Integrated services in the Internet architecture: an overview. *Internet RFC 1633*, June 1994.
13. Wroclawski J. Specification of the controlled-load network element service. *Internet RFC 2211*, Sept. 1997.
14. Shenker S, Partridge C, Guérin R. Specification of guaranteed quality of service. *Internet RFC 2212*, Sept. 1997.
15. Guérin R, Li L, Nadas S, Pan P, Peris V. The cost of QoS support in edge devices: An experimental study. In *Proceedings of IEEE INFOCOM'99*, Mar. 1999.
16. Heinanen J, Baker F, Weiss W, Wroclawski J. Assured forwarding PHB group. *Internet RFC 2597*, June 1999.
17. Jacobson V, Nichols K, Poduri K. An expedited forwarding PHB. *Internet RFC 2598*, June 1999.
18. Charny A, Baker F, Bennett J, Benson K, Boudec JYL, Chiu A, Courtney W, Davie B, Davari S, Firoiu V, Kalmanek C, Ramakrishnan K, Stiliadis D. Supplemental information for the new definition of the EF PHB. *Internet RFC 3247*, March 2002.
19. Rezende JF. Assured service evaluation. In *Proceedings of the IEEE Global Telecommunications Conference - GLOBECOM'99*, Dec. 1999.
20. Dovrolis C, Ramanathan P. A case for relative differentiation services and the proportional differentiation model. *IEEE Network* 1999; **13**(5):26–34.
21. Semret N, Liao RRF, Campbell AT, Lazar AA. Peering and provisioning of differentiated Internet services. In *Proceedings of IEEE INFOCOM'2000*, Mar. 2000.
22. Hao F, Zegura E, Ammar M. Supporting server selection in differentiated service networks. In *Proceedings of IEEE INFOCOM'2001*, Apr. 2001.
23. Fiedler U, Huang P, Plattner B. Towards provisioning DiffServ intra-nets. In *Proceedings of the 9th IEEE/IFIP International Workshop on Quality of Service - IWQoS'01*, June 2001.
24. Li B, Hamdi M, Jiang D, Cao XR, Hou YT. QoS-enabled voice support in the next-generation Internet: Issues, existing approaches and challenges. *IEEE Communications Magazine* 2000; **38**(4):54–61.
25. Hassan M, Nayandoro A, Atiquzzaman M. Internet telephony: Services, technical challenges, and products. *IEEE Communications Magazine* 2000; **38**(4):96–103.
26. Nichols K, Jacobson V, Zhang L. A two-bit differentiated services architecture for the Internet. *Internet RFC 2638*, July 1999.
27. Katevenis M, Sidiropoulos S, Courcoubetis C. Weighted round-robin cell multiplexing in a general-purpose atm switch chip. *IEEE Journal on Selected Areas in Communications* 1991; **9**(8):1265–1279.
28. Bennett J, Benson K, Charny A, Courtney WF, Boudec JYL. Delay jitter bounds and packet scale rate guarantee for expedited forwarding. In *Proceedings of IEEE INFOCOM'2001*, Apr. 2001.
29. Jacobson V, Nichols K, Poduri K. The virtual wire behavior aggregate. *Internet Draft*, Mar. 2000. `draft-ietf-diffserv-ba-vw-00`.
30. Schulzrinne H. Converging on Internet telephony. *IEEE Internet Computing* 1999; **3**(3):40–43.
31. Goyal P, Greenberg A, Kalmanek CR, Marshall WT, Mishra P, Nortz D, Ramakrishnan KK. Integration of call signaling and resource management for IP telephony. *IEEE Internet Computing* 1999; **3**(3):44–52.
32. Su D, Srivastava J, Yao JH. Investigating factors influencing QoS of Internet phone. In *Proceedings of the IEEE International Conference on Multimedia Computing and Systems – ICMCS'99*, June 1999.
33. ITU-T. One-way transmission time. *Recommendation G.114*, Mar. 1993.
34. Li GL, Li B, Cui JH, Li FM. Transient QoS guarantee for packet voice. In *Proceedings of the IEEE Global Telecommunications Conference - GLOBECOM'98*, Nov. 1998.
35. Naser H, Leon-Garcia A, Aboul-Magd O. Voice over differentiated services. *Internet Draft*, Dec. 1998. `draft-naser-voice-diffserv-eval-00`.
36. Ziviani A, Rezende JF, Duarte OCMB. Towards a differentiated services support for voice traffic. In *Proceedings of the IEEE Global Telecommunications Conference - GLOBECOM'99*, Dec. 1999.
37. Ziviani A, Rezende JF, Duarte OCMB, Fdida S. Evaluating voice traffic in a differentiated services environment. In *Proceedings of the 17th International Teletraffic Congress (ITC17)*, Dec. 2001.
38. Karam MJ, Tobagi FA. Analysis of the delay and jitter of voice traffic over the Internet. In *Proceedings of IEEE INFOCOM'2001*, Apr. 2001.

39. Fall K, Varadhan K. NS – notes and documentation. *Tech. rep.*, The VINT Project, Jan. 1999.
40. Network Simulator (NS). `http://www-mash.cs.berkeley.edu/ns` [16 July 2002].
41. Claffy K, Miller G, Thompson K. The nature of the beast: Recent traffic measurements from an Internet backbone. In *Proceedings of the INET'98*, July 1998.

AUTHORS' BIOGRAPHIES

Artur Ziviani received the B.Sc. degree in Electronics Engineering in 1998 and the M.Sc. degree in Electrical Engineering (emphasis in Computer Networking) in 1999, both from the Federal University of Rio de Janeiro (UFRJ), Brazil. From 1997 to 2000, he was with the Networking and Automation Group (GTA/UFRJ), Rio de Janeiro, Brazil. Since 2000, he is a Ph.D. candidate in Computer Networking under a cooperation between the GTA/UFRJ and the LIP6 (Laboratoire d'Informatique de Paris 6) at the Université Pierre et Marie Curie (Paris 6), Paris, France. His research interests are in quality of service in the Internet and provision of location-aware services.

José F. de Rezende received the B.Sc. and M.Sc. degrees in Electronics Engineering from the Universidade Federal do Rio de Janeiro in 1988 and 1991, respectively. He received the Ph.D. degree in Computer Science from the Université Pierre et Marie Curie, Paris, France, in 1997. He was an associate researcher at LIP6 (Laboratoire d'Informatique de Paris 6) during 1997. Since 1998 he is an Associate Professor at Universidade Federal do Rio de Janeiro (UFRJ). His research interests are in distributed multimedia applications, multipeer communication, high speed networks, and quality of service in the Internet.

Otto Carlos M. B. Duarte was born in Rio de Janeiro, Brazil, on October 23, 1953. He received the Electronic Engineer degree and the M. Sc. degree in Electrical Engineering from Universidade Federal do Rio de Janeiro, Brazil, in 1976 and 1981, respectively, and the Dr. Ing. degree from ENST/Paris, France, in 1985. Since 1978 he is Associate Professor at Universidade Federal do Rio de Janeiro. From January 1992 to June 1993 he has worked at MASI Laboratory in Paris 6 University. In 1995, he has spent three months at International Computer Science Institute (ICSI) associated to the University of California at Berkeley. In 1999 and 2001, he worked as invited professor at Paris 6 University. He is heading the computer network group (Grupo de Teleinformática e Automação - GTA). His major research interests are in multicast, QoS guarantees, security, and mobile communications.