Towards a Differentiated Services Support for Voice Traffic

Artur Ziviani
artur@gta.ufrj.br

José F. de Rezende
rezende@gta.ufrj.br

Otto Carlos M. B. Duarte
otto@gta.ufrj.br

Grupo de Teleinformática e Automação (GTA)
Universidade Federal do Rio de Janeiro
COPPE/EE - Programa de Engenharia Elétrica
Caixa Postal 68504 - 21945-970 - Rio de Janeiro - RJ - Brasil

Abstract

The Differentiated Services architecture offers a scalable alternative to provide Quality of Service to the new multimedia applications in the Internet. This paper studies the transport of voice traffic for the Expedited Forwarding (EF) scheme, which is one of the current proposals in the Differentiated Services context. The services provided to traffic generated by On-Off and CBR sources are compared in terms of end-to-end delay and jitter. The analysis includes the different algorithms that could implement the EF service, the influence of the number of active flows, and how efficiently each type of traffic uses an extra allocated bandwidth.

1 Introduction

The current Internet architecture does not meet the Quality of Service (QoS) requirements of the new multimedia applications [1]. The best-effort service offered by the IP protocol does not provide any guarantees in terms of packet delivery delay or packet loss. Packets are treated equally through their routes in the core of the network. Therefore, packets from applications that could tolerate some level of packet loss, delay or jitter compete for resources in equal conditions with packets from applications that have strong real-time requirements. In addition to these features, the Internet presents an explosive growth in the number of connected stations, in the amount of available information and, as a consequence, in the transported traffic. The new multimedia applications have requirements that are different from those of traditional applications. This situation indicates a need for additional services in the Internet.

The proposal of a new set of services developed at the Internet Engineering Task Force (IETF) is called Internet Integrated Services (IIS) [2]. In an Integrated Services environment, each data flow represents an open session. The Integrated Services architecture recommends the use of RSVP (Resource Reservation Protocol) [3, 4, 5] for signaling the resource reservation requests of sessions. The maintenance of one state per flow and the processing associated with these states are not feasible in routers that have a great volume of traffic in a global scale network as the Internet. This lack of scalability has limited the adoption of the Integrated Services proposal in the Internet.

From the preliminary studies of Integrated Services emerged the proposal of Differentiated Services (DS) [6, 7]. The DS scheme provides scalable discriminated services without the need of one state per flow and signaling at every hop. The treatment offered by a DS compatible node is applied to an aggregation of flows, not more to individual flows. Individual flows are classified in aggregations that are previously set. These aggregations will be differently served within a DS domain. This classification and packet marking happen at a DS domain boundary. Within a DS domain, aggregations are forwarded by each node according to a Per-Hop Behavior (PHB). The PHB is selected through a mapping between a codepoint that identifies the aggregation and this behavior. The end-to-end service received by an aggregation results from the combination of all PHBs along its path within a DS domain.

This paper provides a simulation study of voice traffic transmission over a DS domain. In our simulations, two kind of sources are evaluated. These traffic sources are either a constant bit rate (CBR) source or an On-Off source. Both of them are often used to model voice traffic. Preliminary studies about voice over Differentiated Services are in progress. In [8], voice traffic is transmitted over the original DS structure, where the concept of PHB was not defined. In our paper we focus on the performance of the Expedited Forwarding (EF) PHB [9] when transporting voice traffic modeled by CBR or On-Off sources. The analysis includes the different algorithms that could implement the EF service, the influence of the number of active flows, and how efficiently each type of traffic uses an extra allocated bandwidth.

The rest of the paper is organized as follows: Section 2 reviews some basic concepts and current proposals for Differentiated Services. Section 3 describes the models and the topology used in our simulations. Section 4 presents our simulation results. Finally, we conclude in Section 5.

2 The Expedited Forwarding Scheme

The identification of a flow aggregation within a DS domain is performed by checking a new field called Differentiated Services (DS). The DS field is defined in replacement of the IPv4 TOS octet or the IPv6 Traffic Class octet [10]. This field holds a Differentiated Services Codepoint (DSCP). Each DS capable router maps one codepoint in a Per-Hop Behavior (PHB). This PHB defines a treatment to be offered to the packet until it is forwarded or discarded on each node. Ingress routers of a DS domain must have enough information to identify a flow individually. So that, packets belonging to a flow could be classified and marked with the corresponding codepoint of a specific flow aggregation. To summarize, there is one state per flow at the DS domain boundary and one state per aggregation within the DS domain.

Nowadays, two PHB proposals for implementing Differentiated Services are in discussion: Assured Forwarding (AF) [11] and Expedited Forwarding (EF) [9]. This paper focuses on the second one. The Expedited Forwarding 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 end-to-end service through DS domains. In this paper, we investigate how voice traffic (CBR or On-Off sources) is served by the EF PHB within a DS domain.

Queues along the path are the main reasons for packet loss, latency and jitter of a traffic. The EF PHB intends to reduce the time spent in queues by an EF aggregation 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 independent of the intensity of any other traffic attempting to transit the node. Ingress routers must police the incoming EF traffic. Such a policing mechanism should assure the aggregate's arrival rate at the interior nodes to be compliant with the contracted rate. In addition to protect the DS domain from an abusive use, it forces the EF sources to shape the transmitted traffic if they want to prevent their packets being dropped at the DS boundary.

An EF PHB implementation could be performed by different queueing scheduling mechanisms. A priority queue gives the desired behavior. In this case, the traffic conditioning at the DS boundary becomes specially important. The absence of such conditioning could allow the use of resources that are not contracted, harming traffics in other queues. An alternative is to use a single queue in a group of queues served by a weighted round-robin scheduler. In this case, the share of the output bandwidth assigned to the EF queue must be equal to the configured rate.

In [9], Jacobson et al. analyse how WRR compares to PRR in jitter when the EF traffic is generated by CBR sources. The jitter is defined as the absolute value of the difference between the arrival times of two adjacent packets minus their departure times, as shown in Equation 1. The results demonstrate that a PRR implementation is the best case and the WRR implementation is the worst case when compared in jitter.

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

It is also shown that if the WRR weight is chosen to exactly balance arrival and departure rates the results are not stable. Thus, a service-to-arrival rate ratio is defined. The definition of the service-to-arrival rate ratio is presented in Equation 2. The value WRR_{share} is the percentage of the output link bandwidth (BW_{output}) allocated to the EF aggregation. PR_{EF} is the peak rate of the EF aggregation flow. When we have a service-to-arrival ratio greater than 1, we reserve more bandwidth than it was requested in all links within a DS domain. The minimum service-to-arrival ratio to have stability is 1.03, and the value used in our simulations is set to 1.06, i.e., the actually reserved bandwidth is 6% larger than it has been requested.

$$r = \frac{WRR_{share}.BW_{output}}{PR_{EF}} \tag{2}$$

3 Simulation Model

We implemented our simulation model in a modified ns-2 simulator [12]. 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 used some of such mechanisms from our modified ns-2 to implement the EF PHB.

Classifiers at the DS domain boundary are configured to identify flows belonging to an EF aggregation. The remaining flows are classified into an aggregation served by the best-effort model. Then packets are marked with an appropriate codepoint. Within the DS domain, classifiers select the suitable forwarding scheme for each aggregation. At each node, a queue of relative priority (PRR case) or a service share to be offered to that queue (WRR case) is associated to each codepoint or set of codepoints.

The simulation network topology implemented in ns-2 is shown in Figure 1. It illustrates S_i sources connected to a DS domain by 10Mbps links. Ingress or egress routers are hachured. Interior nodes are interconnected by 2.048Mbps, 10Mbps or 100Mbps links. The domain's bottleneck is a 2.048Mbps link, which represents a E1 long distance point-to-point link. The destination nodes D_i are connected to the DS domain by 10Mbps links. In all links within the DS domain, the reserved bandwidth for the EF aggregation is initially kept at 512kbps. This allocation is equivalent to 25% of the 2.048Mbps link, which is the DS domain's bottleneck. The remaining bandwidth is used by the best-effort traffic.

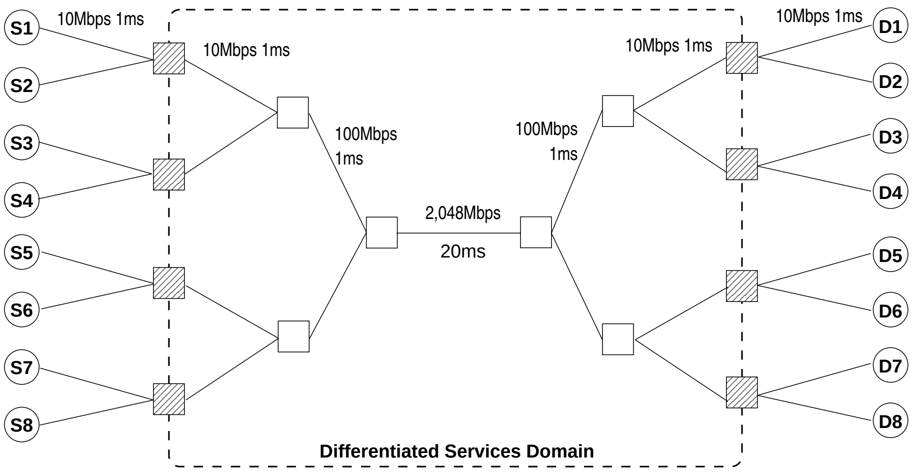


Figure 1: Simulation network topology.

We used CBR or On-Off sources to model voice traffic. When the On-Off source is in the On state, it generates voice traffic at a peak rate of 64kbps. The On periods are determined by a random variable with exponential distribution with mean 400ms. The Off periods are also exponentially distributed by a random variable with mean 600ms. Therefore, the average transmission rate for the On-Off traffic is 25.6kbps. The CBR sources have rates of 64kbps with a variation of ±10% from this value. Such sources are called jittered CBR sources [9]. All flows, CBR or On-Off, have packets of 576 bytes. Sources located at node S_i send packets to node D_i. The sources are always uniformly distributed among the S_i nodes. To fulfill the remaining bandwidth, best-effort traffic is generated by 8 FTP connections, which are active as long as the simulation runs. The size of the FTP packets is 1500 bytes. In a worst case scenario, instead of the 8 FTP connections we use a single CBR source with transmission rate of 1.505Mbps (2.048Mbps - 1.06 $*$ 512kbps), which completely fulfills its share of the bottleneck link.

The traffic generated by the jittered CBR sources or by the On-Off sources are shaped at the output interface of the node where each source is located. This shaping is performed by a leaky bucket mechanism. The depth of each bucket is kept at 576 bytes (1 packet). The token rate varies with the number of EF flows to keep full the 512kbps bandwidth share allocated to the EF aggregation.

4 Results

The results in this section are measured from an EF reference flow whose source is located at node S_1. Its packets are sent to a destination at node D_1. In all simulations, the delay is the end-to-end delay of packets from the reference flow. The jitter is obtained as described in Equation 1. To keep interactivity in a normal conversation, the end-to-end delay of voice traffic in one direction must be below 150ms. A small variation in jitter is also desirable because it reduces the buffer capacity needed in receivers to compensate these variations.

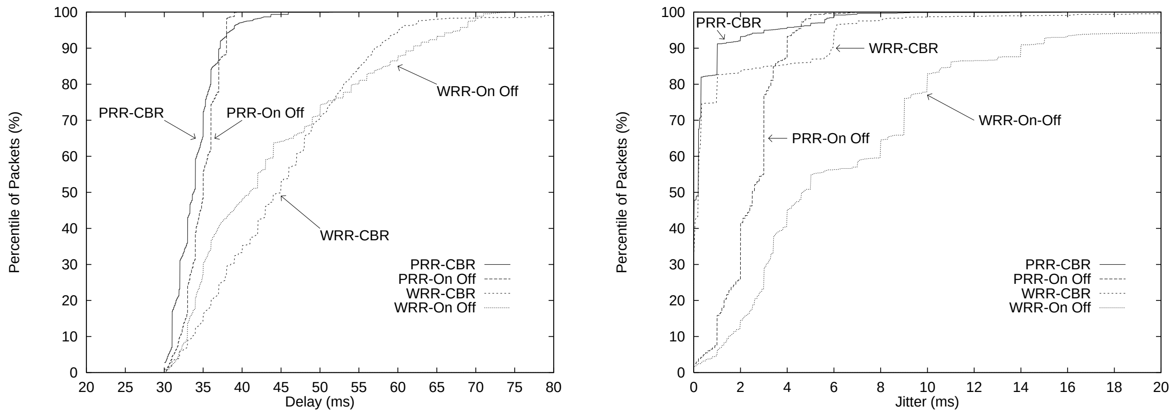
4.1 WRR and PRR Comparison

In Section 2, we pointed out that the EF PHB could be implemented by different queue scheduling mechanisms such as PRR or WRR. Our first set of simulations compares the two implementations when they transport CBR or On-Off traffic as the EF aggregation.

In the CBR case, there are 8 jittered CBR sources (±10% variation) with 64kbps rate. Each such CBR source has its own shaper whose token rate is configured to 64kbps. In the On-Off case, there are 20 sources whose shapers have their token rate set to 25.6kbps, which is the average transmission rate. The service-to-arrival ratio is kept at 1.06. Thus, the actual reserved bandwidth for the EF aggregation is 542.72kbps. To obtain a worst case study, we use a CBR source with a 1.505Mbps rate to constantly fulfill the remaining bandwidth of the bottleneck link.

The end-to-end delay and jitter percentiles of CBR and On-Off traffics in both implementations are shown in Figures 2a and 2b, respectively. The delay introduced by the shapers is not considered because this conditioning is done at the sources and our purpose is to investigate the effects of the DS domain. The PRR implementation of the EF PHB shows a low delay on both CBR and On-Off cases. The delay of all packets served by PRR stands within a narrow range of variation. That is because when there is a packet in the priority queue, it is transmitted harming any other packets in the remaining queues. EF packets only suffer some queueing delay when there are other EF packets ahead of them on the same queue. The jitter of the CBR traffic is quite low on the PRR implementation. The PRR implementation is the best case on both CBR and On-Off cases.

The WRR implementation of the EF PHB shows a higher end-to-end delay. In such implementation, there is an assured share of service to the non-EF traffic. That is, some EF packets may have to wait for the moment when the EF queue will be served. Almost 75% of the On-Off traffic has a lower end-to-end delay than the one presented by the CBR traffic. However, above the 75^{th} delay percentile, the performance of CBR traffic is better than those of On-Off. Such behavior implicates in a higher jitter for the On-Off traffic. The On-Off traffic transported by a WRR implementation has a jitter of 20ms for 90% of the packets and of 21ms for 99% of the packets. If we consider the peak rate of the On-Off traffic, a jitter of 21ms represents a 168 bytes buffer at the receiver. That



(a) Delay (ms)

(b) Jitter (ms)

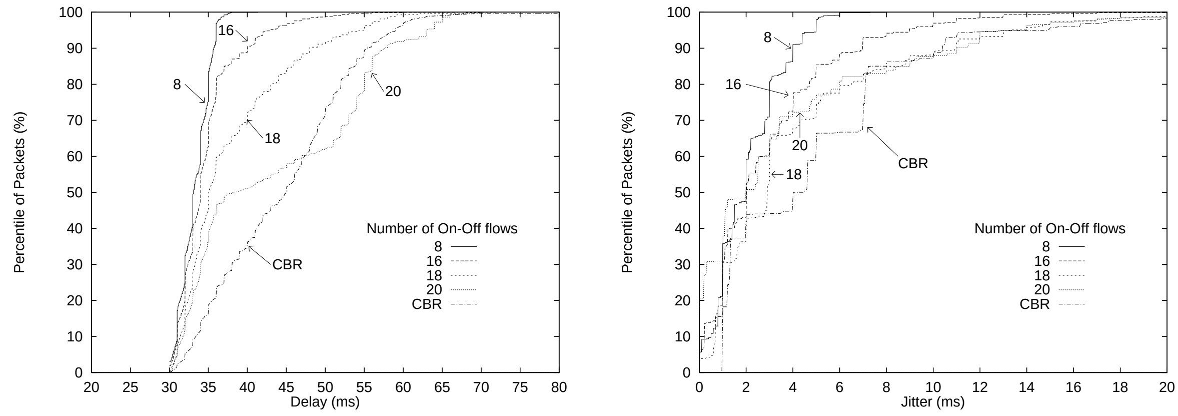
Figure 2: WRR and PRR comparison.

is, a buffer equivalent to a packet (576 bytes) is by far enough. This result makes evident how low is the jitter offered by an Expedited Forwarding service.

The current commercial routers already have an implementation of the WRR queue scheduling algorithm. Because of this, although the WRR implementation has a lower performance than PRR, alternatives that could increase its performance deserve to be investigated. In the next simulations, we verify if changes in some simulation parameters could improve the WRR performance.

4.2 Sensitivity to the Number of On-Off Flows

The next simulation investigates if a variation in the number of active On-Off flows can reduce the end-to-end delay and jitter of this traffic in a WRR implementation of the EF PHB. In this simulation, the non-EF traffic is composed by the 8 FTP connections as described in Section 3. Token rates at shapers are reconfigured in such a way that the EF aggregation keeps a total 512kbps rate, which is the portion of the bottleneck link allocated to the EF aggregation. The number of flows varies from 8 On-Off flows, with an individual token rate of 64kbps, to 20 On-Off flows, with an individual token rate of 25.6kbps. We also analyse if a variation in the number of active On-Off flows leads this



(a) Delay (ms)

(b) Jitter (ms)

Figure 3: Number of On-Off flows variation.

traffic to a better performance when compared to CBR traffic. The end-to-end delay and jitter percentile results are shown in Figures 3a and 3b, respectively. Equivalent results to 8 CBR flows are included as a baseline for comparison.

The results obtained for the 20 On-Off and the 8 CBR flows are better than the ones of Subsection 4.1, since the supplementary traffic is now generated by 8 FTP sources. When a FTP connection loses a packet, it backs off its transmission rate, and increases it gradually until a new loss occurs. Because of this behavior, the occupation of the bottleneck link by the supplementary traffic is not constant. This happens in contrast with our worst case study where we have a 100% occupation of the bottleneck link.

The performance obtained by the On-Off traffic improves as the amount of flows decreases. Note that we vary the token rate at the shapers attached to each source according to the number of active On-Off flows. This is performed in order to keep the allocated 512kbps bandwidth full. With 18 On-Off flows we get a better performance than the one that is reached by 8 CBR flows when both aggregations use the same amount of reserved bandwidth within a DS domain. Therefore, the number of On-Off flows that could be absorbed by a DS domain is considerably larger than the number of CBR flows while obtaining a similar performance.

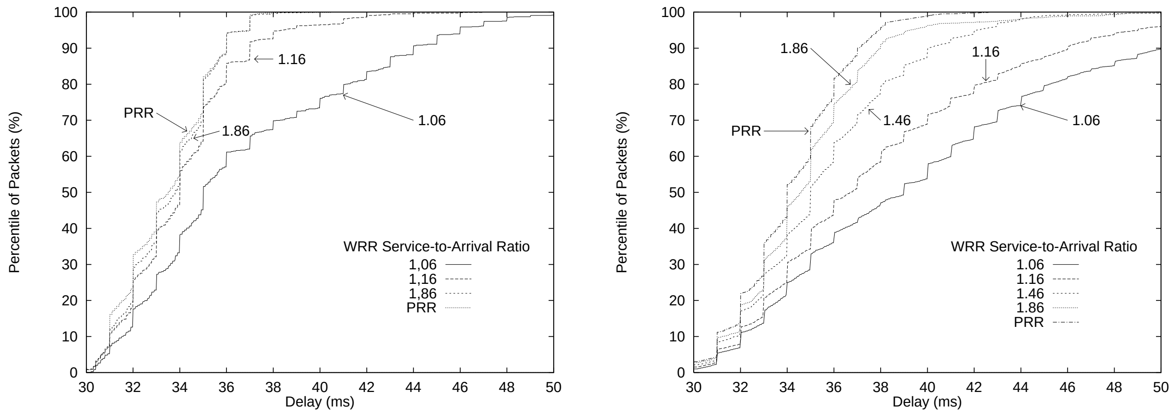
4.3 Service-to-Arrival Ratio Variation

In our third scenario, we study the end-to-end delay variation with the service-to-arrival ratio (Equation 2). The best-effort traffic is also generated by 8 FTP sources, as in Subsection 4.2. The service-to-arrival ratio is varied from 1.06 to 1.86. In the last case, the portion of the bottleneck link allocated to the EF aggregation represents 46.5% of the link or 952kbps. We made two simulations, each one with a different EF aggregation served by a WRR implementation of the EF PHB. In the former, the EF aggregation is composed by 20 On-Off flows shaped to a 25.6kbps rate. In the latter, 8 CBR sources generate the components of the EF aggregation all shaped to 64kbps. The goal of this simulation is to investigate how each kind of traffic uses an extra allocated bandwidth. The end-to-end delays of CBR and On-Off traffics are shown in Table 1 and Figure 4. The performance obtained by the 8 CBR sources when using PRR is shown as a reference.

Table 1: Delay variation with the service-to-arrival ratio.

			Delay Percentile (ms)		
	r	Average Delay (ms)	50	90	95
On-Off	1.06	36.6597	35.0	44.0	46.0
	1.16	34.1081	34.0	37.0	38.3
	1.46	33.5674	33.8	36.0	36.9
	1.86	33.5674	33.8	36.0	36.9
	PRR	33.442	33.5	36.0	37.0
CBR	1.06	39.8603	39.0	50.0	52.8
	1.16	37.9358	36.7	46.0	49.0
	1.46	35.6533	35.0	40.0	42.1
	1.86	34.8192	34.6	38.0	39.3
	PRR	34.2107	34.0	37.0	38.0

As expected, an increase in the service-to-arrival ratio r reduces the end-to-end delay in both traffics. Nevertheless, in delay performance terms, the EF aggregation composed by On-Off flows improves its performance faster than the one composed by CBR flows.



(a) On-Off (b) CBR

Figure 4: Delay variation with the service-to-arrival ratio.

The On-Off traffic using WRR reaches, with a service-to-arrival ratio of 1.46, a similar performance to that obtained when using PRR. From this point on, performance gains leveled off. The CBR traffic performance gets gradually closer to the one achieved when using a PRR mechanism. This performance is still inferior to PRR with $r = 1.86$. In both cases, there is a performance bound equivalent to the one reached by a PRR implementation of the EF PHB. These results indicate that a reservation of extra resources in routers specially benefits an EF aggregation composed by On-Off flows. The performance gain of an EF aggregation formed by On-Off flows shows that such an aggregation could use more efficiently an extra amount of allocated resources.

5 Conclusion

This paper presented a study about an Expedited Forwarding support for voice traffic in a Differentiated Services Internet environment. Voice traffic was modeled by CBR and On-Off sources. We compared the service offered to both traffics considering two different implementations of the EF PHB. One was a priority round-robin mechanism (PRR) and the other a weighted round-robin mechanism (WRR). The analysis was based on the end-to-end delay and jitter. The use of shapers to condition the traffic at the output interfaces of the sources was fundamental to keep the delay and jitter within an acceptable bound.

We showed that the EF service could support more On-Off flows than CBR flows with similar performance in terms of jitter. The results indicate that an extra reserved bandwidth is more efficiently used by On-Off flows than by CBR flows. So that, for the same extra bandwidth, the performance gain is larger on On-Off traffics. However, this improvement in performance is bounded by the performance of a PRR implementation. The study of alternatives to improve the performance of a WRR implementation is justified when we know that this algorithm is already available on commercial routers. In such case, an adequate configuration of today's routers can provide the proposed behavior of Expedited Forwarding. This paper contributes with guidelines to administrators that, in the future, may want to configure their routers to offer a Differentiated Services support for voice traffic through the EF scheme.

We pointed out the importance of traffic conditioning in the context of a Expedited Forwarding implementation. The effects of this conditioning over On-Off or CBR sources deserves to be further investigated. A variation on the bucket depth of shapers, allowing some bursty traffic into the DS domain, implicates in changes on the performance of flows. As future work, we plan to investigate the delay introduced by traffic shapers and the consequence on the On-Off or CBR traffic of a variation in the bucket depth of the shapers.

References

[1] C. Lefelhocz, B. Lyles, S. Shenker and L. Zhang, "Congestion control for best-effort service: Why we need a new paradigm", *IEEE Network*, vol. 10, no. 1, pp. 10–19, Jan. 1996.

[2] R. Braden, D. D. Clark and S. Shenker, "Integrated services in the Internet archi-

tecture: an overview", *Internet RFC 1633*, June 1994.

[3] L. Zhang, S. E. Deering, D. Estrin, S. Shenker and D. Zappala, "RSVP: A new resource reservation protocol", *IEEE Network*, vol. 7, no. 5, pp. 8–18, Sept. 1993.

[4] R. Braden, L. Zhang, S. Berson, S. Herzog and S. Jamin, "Resource reservation protocol (RSVP) – version 1 functional specification", *Internet RFC 2205*, Sept. 1997.

[5] P. P. White, "RSVP and integrated services in the Internet: A tutorial", *IEEE Communications Magazine*, vol. 35, no. 5, pp. 100–106, May 1997.

[6] S. Blake, D. L. Black, M. Carlson, E. Davies, Z. Wang and W. Weiss, "An architecture for differentiated services", *Internet RFC 2475*, Dec. 1998.

[7] F. Baumgartner, T. Braun and P. Habergger, "Differentiated services: A new approach for quality of service in the Internet", in *Proceedings of the 8th IFIP Conference on High Performance Networking – HPN'98*, Sept. 1998.

[8] H. Naser, A. Leon-Garcia and O. Aboul-Magd, "Voice over differentiated services", *Internet Draft*, Dec. 1998. `draft-naser-voice-diffserv-eval-00`.

[9] V. Jacobson, K. Nichols and K. Poduri, "An expedited forwarding PHB", *Internet Draft*, Feb. 1999. `draft-ietf-diffserv-phb-ef-02`.

[10] K. Nichols, S. Blake, F. Baker and D. L. Black, "Definition of the differentiated services field (DS field) in the IPv4 and IPv6 headers", *Internet RFC 2474*, Dec. 1998.

[11] J. Heinanen, F. Baker, W. Weiss and J. Wroclawski, "Assured forwarding PHB group", *Internet Draft*, Feb. 1999. `draft-ietf-diffserv-af-06`.

[12] Network Simulator - NS (version 2). `http://www-mash.cs.berkeley.edu/ns/`.