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 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. The best-effort service offered by the IP protocol does not provide any guarantees of packet delivery delay or packet loss. 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. Besides, 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. 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) [1]. Nevertheless, its lack of scalability limits the widespread adoption of the IIS proposal in the Internet.

From the studies of Integrated Services emerged the proposal of Differentiated Services (DS) [2]. 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-compliant node is applied to an aggregation of flows, not more to individual flows. Individual flows are classified in previously set aggregations. These aggregations will be differently served within a DS domain. The 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 service received by an aggregation within a DS domain results from the combination of all PHBs along its path.

This paper provides a simulation study of voice traffic transmission over a DS domain. In our simulations, the voice traffic sources are either modeled by constant bit rate (CBR) sources or by On-Off sources. In [3], 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 [4] when transporting voice traffic. 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 voice 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, pointing out the Expedited Forwarding scheme. 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 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 [5]. This field holds a Differentiated Services Codepoint (DSCP). Each DS capable router maps one codepoint in a Per-Hop Behavior (PHB). The 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 identify a flow individually. That is, 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.

Currently, two PHB proposals for implementing Differentiated Services are in discussion: Assured Forwarding (AF) [6] and Expedited Forwarding (EF) [4]. 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 may be used to build a low loss, low latency, low jitter, assured bandwidth service within 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 independently 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.

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

Jacobson *et al.* [4] analyze how WRR compares with 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, then

$$jitter = |(a_j - a_i) - (d_j - d_i)| \quad (1)$$

The results demonstrate that a PRR implementation is the best case and the WRR implementation is the worst case when compared in jitter. 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 r is defined as

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

where 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.

3 Simulation Model

We implemented our simulation model in a modified ns-2 simulator [7]. 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.

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

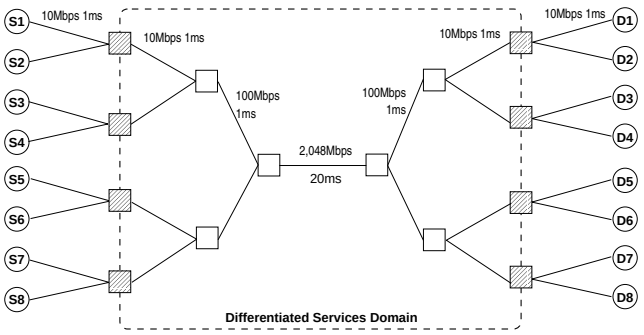


Figure 1: Simulation network topology.

We used CBR or On-Off sources to model voice traffic. When an On-Off source is in the On state, it generates voice traffic at 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. The CBR sources have rates of 64 kbps with a variation of $\pm 10\%$ from this value. Such sources are called jittered CBR sources [4]. All flows, CBR or On-Off, have packets of 576 bytes. Sources located at node S_i send packets to node D_i. To fulfill the remaining bandwidth, best-effort traffic is generated by 8 FTP connections that are active as long as the simulation runs. The size of the FTP packets is 1500 bytes. In a high load scenario, instead of the 8 FTP connections we use a single CBR source with transmission rate of 1.505 Mbps (2.048 Mbps - $1.06 * 512$ kbps) that 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 token 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 512 kbps 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. 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 150 ms [8]. A small variation in jitter is also desirable because it reduces the buffer capacity needed in receivers to compensate these variations.

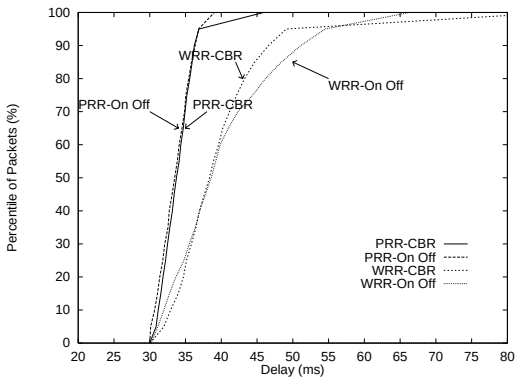
4.1 WRR and PRR Comparison

Our first set of simulations compares the PRR or WRR implementations when they transport CBR or On-Off traffic as the EF aggregation. In the CBR case, there are 8 jittered CBR sources ($\pm 10\%$ variation) with 64 kbps rate. Each such CBR source has its own shaper whose token rate is configured to 64 kbps. In the On-Off case, there are 20 sources whose shapers have their token rate set to 25.6 kbps, which is the average transmission rate. The service-to-arrival ratio r is kept at 1.06. Thus, the actual reserved bandwidth for the EF aggregation is 542.72 kbps. To obtain a high load study, we use a CBR source with a 1.505 Mbps rate to constantly fulfill the remaining bandwidth of the bottleneck link. The 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.

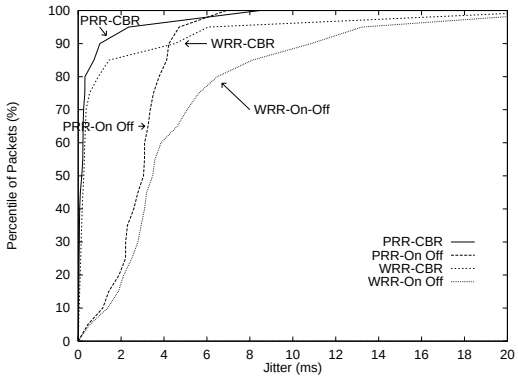
The WRR implementation of the EF PHB shows a higher delay. In such implementation, there is an assured share of service to the non-EF traffic. That is, some EF packets may have to wait for the moment when the EF queue will be served.

4.2 Sensitivity to the Number of On-Off Flows

The next simulation investigates how much a variation in the number of active On-Off flows can reduce the 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. The delay and jitter percentile results are shown in Figures 3a and 3b, respectively. Equivalent

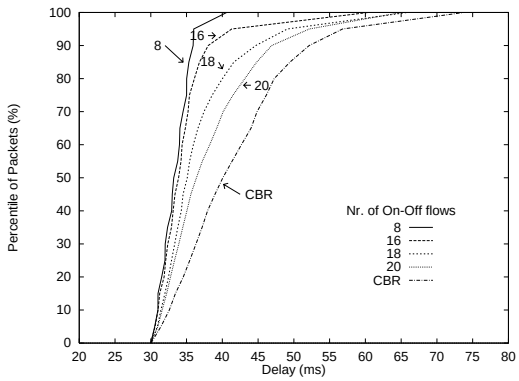


(a) Delay (ms)

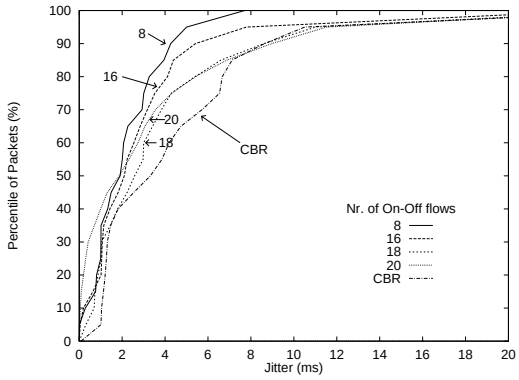


(b) Jitter (ms)

Figure 2: WRR and PRR comparison.



(a) Delay (ms)



(b) Jitter (ms)

Figure 3: Number of On-Off flows variation.

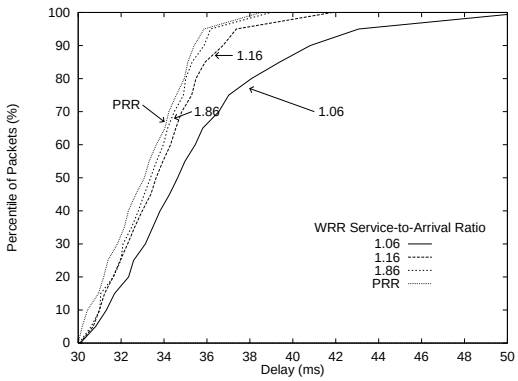
results to 8 CBR flows are included as a baseline for comparison.

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 high load study where we have a 100% occupation of the bottleneck link.

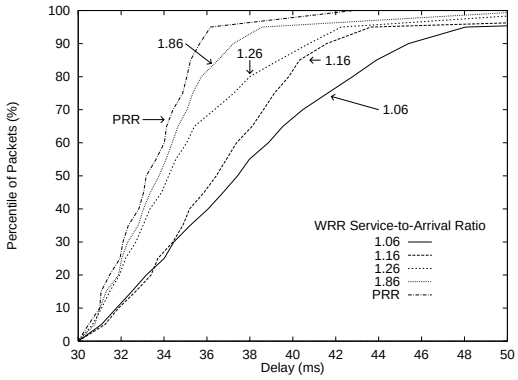
The performance obtained by the On-Off traffic improves as the amount of flows decreases. Note that we vary the token rate at the shapers attached to each source according to the number of active On-Off flows. Such variation is performed in order to keep the allocated 512 kbps bandwidth full.

4.3 Service-to-Arrival Ratio Variation

In our third scenario, we study the delay variation with the service-to-arrival ratio r (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. The EF aggregation is composed by 20 On-Off flows shaped to a 25.6 kbps rate. The 8 jittered CBR sources generate the components of the EF aggregation all shaped to 64 kbps. The goal of this simulation is to investigate how each kind of traffic uses an extra allocated bandwidth. The delays of CBR and On-Off traffics are shown in Figure 4. The performance obtained by the 8 CBR sources when using PRR is shown as a reference.



(a) On-Off Traffic



(b) CBR Traffic

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

As expected, an increase in the service-to-arrival ratio r reduces the 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. 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. 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 environment. The study of alternatives to improve the performance of a WRR implementation of EF service 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. 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] R. Braden, D. D. Clark and S. Shenker, "Integrated services in the Internet architecture: an overview", *Internet RFC 1633*, June 1994.

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

[3] H. Naser, A. Leon-Garcia and O. Aboul-Magd, "Voice over differentiated services", *Internet Draft*, Dec. 1998.

[4] V. Jacobson, K. Nichols and K. Poduri, "An expedited forwarding PHB", *Internet RFC 2598*, June 1999.

[5] 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.

[6] J. Heinanen, F. Baker, W. Weiss and J. Wroclawski, "Assured forwarding PHB group", *Internet RFC 2597*, June 1999.

[7] K. Fall and K. Varadhan, "NS – notes and documentation", tech. rep., The VINT Project, May 1998.

[8] ITU-T, "One-way trasmission time". Recommendation G.114, Mar. 1993.