Endpoint Admission Control with Delay Variation Measurements for QoS in IP Networks

G. Bianchi
Universitá di Palermo
Viale delle Scienze
90128 Palermo, ITALY
bianchi@elet.polimi.it

F. Borgonovo, A. Capone, L. Fratta
Politecnico di Milano
Piazza Leonardo da Vinci 32
20133 Milano, ITALY
capone@elet.polimi.it

C. Petrioli
Rome University "La Sapienza"
via Salaria, 113
00198 Roma, ITALY
petrioli@dsi.uniroma1.it

ABSTRACT

In this paper we describe a novel Endpoint Admission Control scheme (EAC) for IP telephony. EAC mechanisms are driven by independent measurements taken by the edge nodes on a flow of packets injected in the network to probe the source to destination path. Our scheme is characterized by two fundamental features. First, it does not rely on any additional procedure in internal network routers other than the capability to apply different service priorities to probing and data packets. Second, the connection admission decision is based on the analysis of the probing flow delay variation statistics. Simulation results, which focus on a IP telephony scenario, show that, despite the lack of core routers cooperation, toll-quality performance figures (99th delay percentiles not greater than few ms per router) can be obtained even in severe overload conditions. Finally, a comparison with an EAC scheme driven by probe losses only, shows that the use of delay variation statistics as endpoint decision criterion is a key factor for EAC effectiveness.

Keywords

Quality of Service, Admission Control, DiffServ, IP

1. INTRODUCTION

It is widely accepted that the today best effort Internet is not able to satisfactorily support emerging services and market demands, such as IP Telephony. Real-time services, in general, and IP telephony, in particular, have very stringent delay and loss requirements (less than 150 ms mouth-to-ear delay for toll quality voice), that need to be met over the whole call holding time. The analysis of the delay components over the source-to-destination path shows that up to 100-150 ms can be spared for compression, packetization, jitter compensation, propagation delay, etc [1], leaving no more than few tens of ms for queueing delay within the many routers on the path.

Many different proposals aimed at achieving such a tight QoS control on the Internet have been discussed in IETF. IntServ/RSVP (Resource reSerVation Protocol) [2, 3] provide end-to-end per-flow QoS by means of hop-by-hop resource reservation within the IP network. Such an approach imposes a significant burden on the core routers, which are required to handle per flow signaling, to maintain per flow forwarding information on the control path, and to perform per flow admission control, classification and scheduling.

To reduce the complexity within each core router, alternative schemes, referred to as Measurement Based Admission Control (MBAC), have been proposed [4, 5, 6, 7]. These schemes replace per-flow states with run-time link load estimates performed in each router. However, MBAC solutions still require significant modification of the existing Internet architecture, as core routers must support load estimation algorithms, and still need to be explicitly involved in per flow signaling exchange.

A completely different approach is provided by Differentiated Services (DiffServ) [8, 9]. In DiffServ, core routers are stateless and unaware of any signaling. They merely implement a suite of buffering and scheduling techniques and apply them to a limited number of traffic classes, whose packets are identified on the basis of the DS field in the IP packet header. As a result, a variety of services can be constructed by a combination of: (i) setting packets DS bits at network boundaries, (ii) using those bits to determine how packets are forwarded by the core routers, and (iii) conditioning the marked packets at network boundaries in accordance with the requirements or rules of each service.

While DiffServ easily enable resource provisioning performed on a management plane for permanent connections, their widely recognized limit is the lack of support for per-flow resource management and admission control, resulting in the lack of strict per flow QoS guarantees. A number of proposals, presented in the literature, have shown that per flow Distributed Admission Control schemes can be deployed over a DiffServ architecture [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]. Although significantly different in implementation details, these proposals, referred hereafter as Endpoint Admission Control (EAC) (according to the overview paper [20]), share the common idea that accept/reject decisions are taken by the network endpoints and are based on the pro-

cessing of "probing" packets, injected in the network at set-up to verify the network congestion status. This approach is a radical overhaul of existing admission control schemes, in which all the routers in the source to destination path are involved in call admission and make an accept/reject decision based on their occupancy status.

Besides their common approach to the problem, EAC schemes show remarkable differences. Some of them rely on some level of internal network routers cooperation (e.g. probing packet marking [19, 20], and ad-hoc probing packet management techniques [18]), while other EAC schemes [11, 12, 13], hereafter referred to as "pure EAC", only require features already available in current routers, e.g. the capability of distinguishing between probing and data packets (e.g. via TOS precedence bits, or the DiffServ DSCP field), and of configuring elementary buffering and scheduling schemes. In particular, the only router requirement in [12] is a priority-based forwarding scheme applied to probing and data packets, while in [11] a strict limit on the probing buffer size is also enforced. Therefore, an important advantage of "pure" EAC solutions, is that they can be easily adopted in the present Internet with no impact on the existing core routers and network infrastructure.

In this paper, we propose a new "pure" EAC scheme, called PCP-DV (Phantom Circuit Protocol- Delay Variation), which is an improved version of the EAC scheme presented in [13]. PCP-DV bases the acceptance test on probing packets delay variation analysis. The major innovative contribution of the paper is twofold. First we discuss why a decision criterion based on probing packets delay variation analysis results into an effective way to control the network congestion status, and we provide a thorough performance evaluation in an IP telephony scenario, for a wide range of parameter settings, which shows that PCP-DV is indeed capable of providing 99th delay percentiles not greater than few ms per router even in heavy overload conditions. Second, by means of a thorough comparison of PCP-DV performance with other alternative "pure" EAC schemes [11], we prove that EAC schemes based on delay variation analysis achieve a more effective link load control, and therefore QoS guarantees, than those based on loss measurements.

The paper is organized as follows. In section 2, PCP-DV operation is described, and the decision criterion rationale is provided. Section 3 describes the simulation model, and presents the VoIP (Voice over IP) variable bit rate traffic scenario adopted to evaluate the protocol performance. Section 4 is dedicated to PCP-DV performance evaluation and parameters tuning, while section 5 compares PCP-DV performance with that of other pure EAC schemes based on probing packet losses. Finally, conclusions are drawn in section 6.

2. PCP-DV OPERATION

In PCP-DV, a user that wants to setup a connection starts a preliminary *Probing Phase* aimed at verifying whether there are enough resources in the network to accept a new connection. The PCP-DV probing phase, graphically shown in Figure 1, consists in the consecutive transmission of N_p probing packets with fixed inter-departure time I. Packets transmitted during the probing phase are marked with

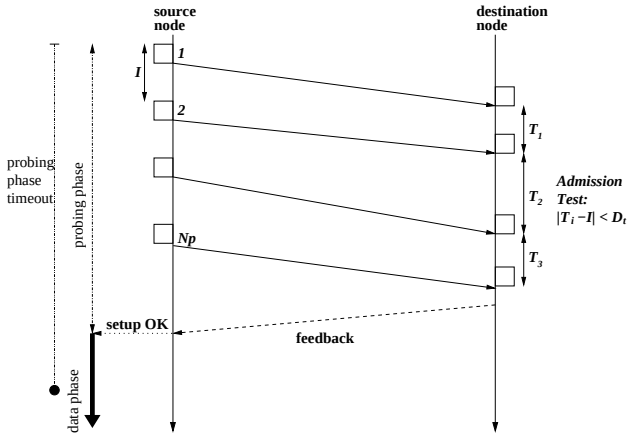

Figure 1: PCP probing and data phases

a different label (a different value in the TOS precedence or DSCP field of the IP header) and are served at core routers with lower priority than data packets. The decision on whether to admit or reject a new call is taken at the destination node, based on jitter measures on the probing flow arrival statistics. Upon reception of the first probing packet, the destination node starts a timer to measure the time T to the next probing packet arrival. If condition

$$I - D_t \leq T \leq I + D_t \tag{1}$$

is met, the timer is restarted and the above procedure is iterated until all N_p packets are received. Conversely, if condition (1) fails for one received probing packet, the connection is rejected. D_t is a parameter that represents the maximum tolerance on the received probing packets jitter. The parameters D_t and N_p regulate the PCP-DV admission mechanism behavior and, as shown in section 2.1, they allow to tune the accepted traffic load and the quality provided to accepted connections.

The final result of the acceptance test is notified back to the source node by means of one or more *Feedback Packets* which can be forwarded with high priority as their contribution to traffic is negligible. In particular, if all the N_p probing packets are received and the acceptance test (1) is always positive, the destination node sends an "accept" feedback packet to the source (for higher reliability, more feedback packets may be sent). Otherwise, if condition (1) is not met for one received probing packet, a "reject" feedback packet is immediately sent back to the source. Upon reception of an "accept" feedback packet, the source node enters a *Data phase* in which information packets are transmitted at high priority according to the traffic source characteristics. If instead a "reject" feedback packet is received, or no feedback arrives at the source node before the *probing phase timeout* expiration, the call is terminated.

The only requirement PCP-DV imposes on the core network is therefore the capability of distinguishing between probe and data/feedback packets (by reading the TOS or DSCP

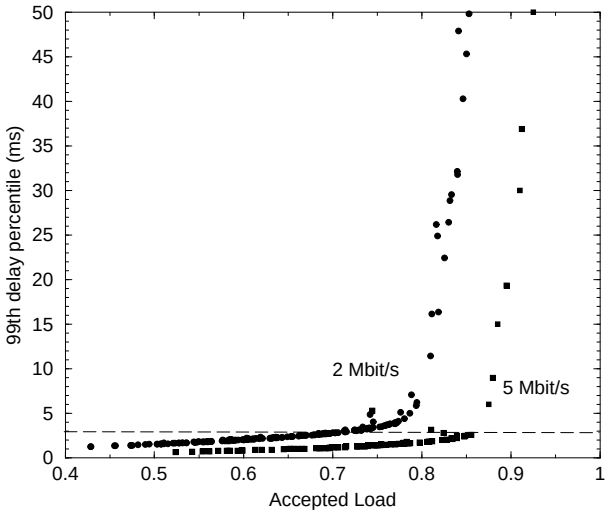

Figure 2: Throughput/delay tradeoffs

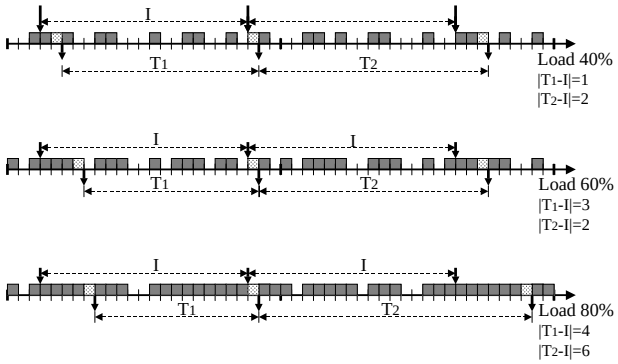

Figure 3: Sensitivity of Probing packets jitter versus accepted load: a probing flow with constant interdeparture probing packets time I is offered to a single link network in three different load conditions. Fixed length packets have been assumed. A black (white) square indicates that a data (probing) packet is being served in the current slot.

field) and of forwarding packets according to a head of the line priority scheme: high priority to data packets, low priority to probing packets. This forwarding mechanism serves a probing packet only when the data packets queue is empty. Probing traffic, which is not admission-controlled, is therefore not allowed to contend bandwidth resources against established traffic, thus preventing data packets QoS degradation. Moreover, since probing packets use only resources unused by accepted calls, the probing packets flow received at the destination contains indirect information on the links congestion status, information that can be used to perform the accept/reject test.

2.1 PCP-DV Rationale

To provide toll quality delay performance, a tight control of the links accepted load is required. Figure 2 reports simulation results quantifying the load/delay relationship in an IP Telephony traffic scenario (the voice traffic model adopted is the Brady ON-OFF model described in section 3). Figure 2 shows the 99th delay percentile of the accepted traffic as function of the accepted traffic load (normalized to the channel capacity and source activity factor), in a 2 and 5 Mb/s single link network. Simulation results have been obtained by running several EAC schemes (either PCP-DV, and the approaches presented in [17, 16]). The results obtained are independent of the specific EAC scheme and related parameter settings adopted, and show that, once a channel capacity is selected, delay performance depends only on the accepted load. By analyzing the two curves shown in Figure 2, we can identify a threshold on the accepted load corresponding to a given delay bound. As an example, a 99th delay percentile equal to few $(3-5)$ ms can be guaranteed by not exceeding an accepted load threshold of $0.74 \div 0.79$ for a 2 Mb/s link, and $0.85 \div 0.87$ for a 5 Mb/s link. The problem of providing quality of service guarantees therefore translates into strictly controlling the link load over the source to destination path.

With EAC schemes links load cannot be directly measured and, therefore, it is necessary to estimate the source-to destination path congestion status based on indirect measures taken at the network endpoints on the probing packets arrival statistics. The PCP-DV acceptance test is based on the observation that a non negligible probing packet jitter can be measured even when the accepted load is well below the channel capacity (e.g. at the target 0.75 normalized accepted load or below), and that probing packets delay variations are very sensitive to link load.

The effect of data traffic load on probing packets jitter is graphically illustrated in the intuitive example shown in figure 3, and quantitatively analyzed in figure 4. The example shows that, owing to the priority-based forwarding scheme, a probing packet is transmitted only when no data packet is stored in the data packet buffer. Therefore, the delay experienced by a probing packet arriving to the link is given by the remaining busy period of the data packets queue. As the accepted load increases, busy periods get longer, and this in turn increases probing packets delay variations, as probing packets arrival times are independent of busy period starting and ending points.

Figure 4 shows the probability, obtained by simulation, that the delay jitter of a single pair of probing packets exceeds a given threshold D_t, versus the accepted load. Four acceptance delay thresholds, respectively equal to 3, 5, 7 and 10 ms have been considered. From the figure, we see that the probability of exceeding a given delay threshold increases with the link load, and gets close to 1 as the link load approaches 1. However, even adopting a small (3 ms) delay threshold, the probability that the jitter is below the threshold (i.e. that test (1) succeeds) is quite high (0.4) at the critical load 0.75.

To improve the power of the acceptance test (1), one must consider several probing packets. This is what proposed in PCP-DV, where N_p-1 acceptance tests need to be successful to ultimately accept the call. Figure 5 shows the probability that a call is rejected versus the link load for different N_p, and D_t equal to 3 and 10 ms.

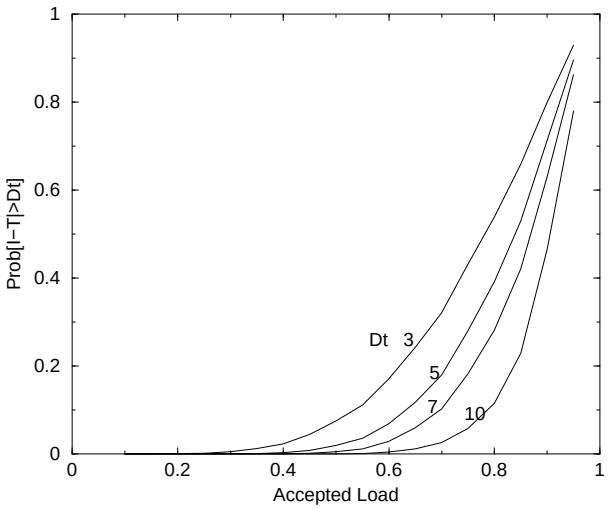

Figure 4: Probability that the probing packets jitter exceeds a given delay threshold (3, 5, 7, and 10 ms), versus link load - 2 Mb/s link capacity

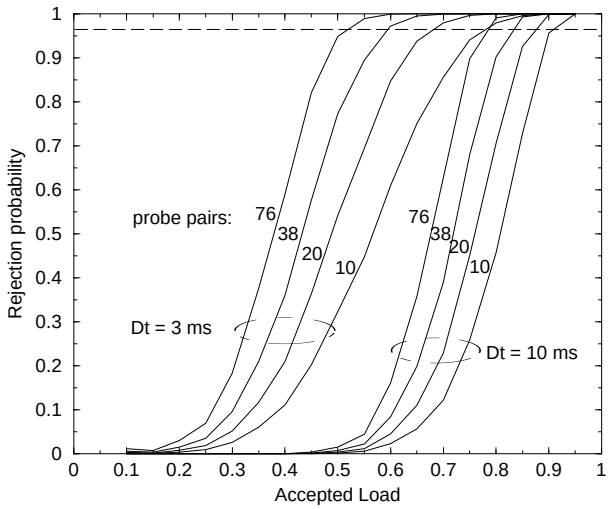

Figure 5: Call rejection probability versus load, for 3 ms and 10 ms delay thresholds, and for a different number of probing packets

¿From the plots in Figure 5 it turns out that the desired high rejection probability at a given threshold traffic load is reached by several parameters settings. However, the slope of the rejection probability curve increases with N_p. Therefore, the effectiveness of the test also increases with N_p as new calls have a very high acceptance probability up to an accepted load close to the threshold. As an example, if the threshold is equal to 0.75, a rejection probability of 0.96 is reached either by $N_p = 11$ and $D_t = 3$ ms or $N_p = 77$ and $D_t = 10$ ms. However, a big difference in the acceptance probability at 0.6 accepted load is observed in the two cases. This behavior suggests to use an N_p as high as possible subject to constraints on the set-up phase maximum length.

2.2 Remarks

In addition to the above described PCP-DV basic principles, the following technical details and remarks are relevant for a thorough comprehension of its operation. The reader is also encouraged to refer to [20] for additional engineering considerations which also apply to PCP-DV.

A. Data Traffic Conditioning

The correct PCP-DV operation, requires that the links load actually reflects the accepted calls in progress. Therefore, in case of VBR traffic, conditioning mechanisms must be adopted. Such conditioning procedures, common to resource reservation techniques based on traffic measurements (see for example [5]), constitute the price to pay for the reduced complexity of the call admission procedure.

B. QoS Support

PCP-DV, differently from stateful centralized solutions, can only provide a single level of QoS. In fact, as long as only two priority levels (probing/data) are used within the network routers, heterogeneous real-time connections, with different loss/delay requirement are forced to share the same data packets queue, and thus, regardless of how sophisticated the end-to-end measurement scheme might be, they ultimately encounter the same performance. To overcome this limitation we envision the same approach described in [20] for EAC schemes. PCP-DV can be used to perform call admission within a DiffServ class. Isolation between DiffServ classes can then be achieved by adopting a WFQ-like mechanism assuring a given rate to the admission-controlled class. However, for the protocol to correctly operate, this mechanism must prevent probing traffic from borrowing bandwidth from other classes, as this may result in call misacceptances.

C. Routing

In all EAC schemes the resources estimation is performed by means of probing packets and if the connection is accepted the corresponding flow is routed on a single path to the destination. If the network routing changes, congestion may be experienced by some flows and the QoS may drop below a desired threshold. However, in the present IP networks, such route changes are usually triggered by a topological change in the network and this kind of events cannot be fought by any resource reservation protocol.

D. Stability

An important feature of PCP-DV is its intrinsic stability and robustness. Indeed, when an increase in the accepted traffic above the QoS limits occurs, e.g. because of rerouting of accepted connections, or misacceptance of calls due to a test failure, thanks to the priority-based forwarding mechanism employed, the probing traffic is throttled. This will prevent acceptance of new connections and, as some calls terminate,

the congestion period will end.

3. SIMULATION MODEL

To evaluate PCP-DV throughput and delay performance, we have used a simulator written in C++. Unless otherwise specified, simulations have been carried out considering a 2 Mb/s single link network scenario. Even if this is a very simplified scenario, it is representative of more complex network scenarios where a bottleneck link exists in the source to destination path, and it is therefore sufficient to investigate PCP-DV behavior.

We have considered an IP telephony variable bit rate traffic where voice sources with silence suppression have been modeled according to the two states (ON/OFF) Brady model [21]. In particular, each voice call alternates between ON and OFF states. During the ON state, the voice source emits vocal talkspurts at a fixed peak rate $B_p = 32$ kb/s, while in the OFF state it is silent. Both ON and OFF periods are exponentially distributed with mean values equal to 1 s and 1.35 s, respectively. The activity factor α is the fraction of time a voice source is found in the ON state. Voice sources are homogeneous, and generate 1000 bit fixed-size packets, corresponding to an inter-departure packet time equal to 31.25 ms when the source is in the ON state.

We have considered a dynamic link load scenario where calls are generated according to a Poisson process and have an exponentially distributed duration. The normalized offered load, ρ, is then defined as:

$$\rho = \frac{\lambda}{\mu} \cdot \frac{B_p \alpha}{C}$$

where λ (calls/s) is the call generation rate, $1/\mu$ (s) is the average call duration, and C is the channel rate (kb/s). Unless otherwise specified, in our simulations, we have adopted $1/\mu = 3$ minutes, the value generally used in telephony.

Upon arrival of a new call, the PCP-DV probing phase has been simulated to determine whether to accept or reject the call. Probing packets are generated at constant rate with packet inter-departure time $I = 26$ ms, which corresponds to a rate 20% higher than the voice peak rate.

Finally, in our simulator we have assumed no loss of feedback packets, and instantaneous feedback packet reception.

4. PERFORMANCE EVALUATION

An extensive performance evaluation has been carried out, by means of simulations, to investigate PCP-DV performance in several network scenarios and to provide insights on PCP-DV parameters tuning. Results are summarized in Fig. 6– Fig. 8. PCP-DV performance has been reported in terms of accepted load and 99-th packet delay percentiles.

Figure 6 shows the normalized accepted load versus the offered load, for link capacity C equal to 2 and 5 Mb/s. PCP-DV parameters N_p and D_t have been selected to guarantee a target accepted load approximately equal to 0.75 for a 2 Mb/s channel (0.85 for a 5 Mb/s channel) when the offered load is 4 times the channel capacity. As discussed in section 2.1, such a target accepted load allows to achieve toll-quality delay performance. The 99th delay percentiles measured in

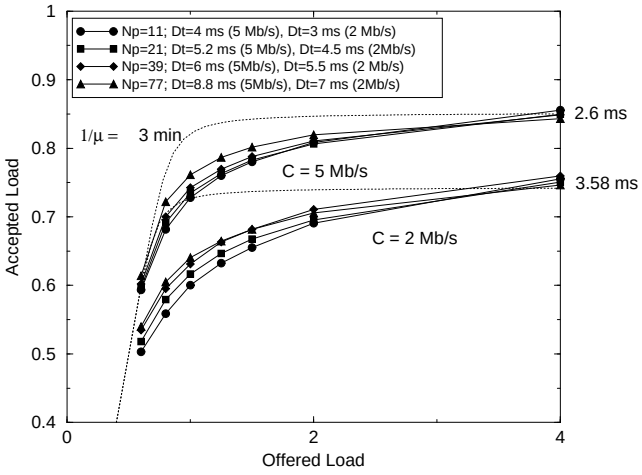

Figure 6: PCP-DV: Accepted vs. Offered load for several N_p values and for 2 and 5 Mb/s link capacity. 99th delay percentiles are also reported for offered load equal to 4.

the simulation at offered load 4 are also reported in the figure (2.6 ms and 3.58 ms for the 2 Mb/s and 5 Mb/s cases, respectively). These results show PCP-DV ability to guarantee very few ms 99-th delay percentile even in very high overload conditions. Figure 6 also shows that different N_p and D_t parameter settings result in similar performance, and, in particular, allow to meet the target accepted load and delay figures at 400 % overload. N_p and D_t have indeed a complementary effect on performance: the shorter the probing phase (i.e. the lower the number of probes N_p), the tighter the delay threshold D_t must be to achieve the target performance.

However, parameters settings achieving the target delay performance in strong overload conditions show different behaviors for offered traffic loads ranging between 0.5 and 1.5. In these practical operational conditions, better performance is obtained, as anticipated in Figure 5, by adopting a longer probing phase ($N_p = 77$) and larger D_t (7 ms). With these parameter values, the acceptance test is less likely to reject calls in underload traffic conditions. This effect can be appreciated in Fig. 6: in the 2 Mb/s case a 10% increase in the accepted load is achieved by the parameters setting $N_p = 77$, $D_t = 7$ms at offered load 1 over the $N_p = 11$, $D_t = 4$ms setting. To optimize the performance it is therefore suggested to choose the longest possible probing phase which allows to meet the constraint on the maximum call setup length (1 s for toll quality IP Telephony).

To quantify the throughput degradation due to unnecessarily rejected calls, intrinsic in any EAC scheme, we have shown in Figure 6 (dotted lines) the performance of an ideal state-full CAC scheme. This upper bound is given by the Erlang-B formula. The degradation of our scheme with respect to the bound decreases as the channel rate increases. We note also that at low load (below 0.6) the degradation is negligible and the difference increases up to 15% in operation conditions which are undesiderable since the call rejection rate is too high ($> 10\%$) even with an ideal CAC.

N_p	$1/\mu = 3$ min	$1/\mu = 10$ min.
11	3.0 ms	2.7 ms
21	4.5 ms	4.2 ms
39	5.5 ms	5.3 ms
77	7.0 ms	6.8 ms

Table 1: Optimal delay threshold, D_t, for call duration of 3 and 10 min

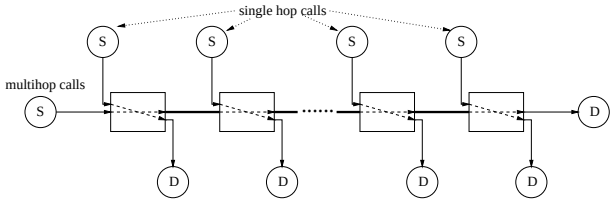

Figure 7: Multi-link network scenario.

To test the robustness of PCP-DV parameter settings we have run simulations with average call duration equal to 10 minutes. At the same offered load, increasing the average call length from 3 to 10 minutes, reduces the arrival rate λ, and thus also the probing load, by a 10/3 factor. Results, reported in Table 1, show that the optimal PCP-DV delay threshold settings is almost the same as in the 3 minutes case. This in turn show that the probing traffic load, and, therefore, the low priority queue status do not significantly impact on PCP-DV acceptance/rejection probability, which only depends on the accepted traffic load (we'll see in section 5 that this property does not hold for mechanisms driven by probing packet losses).

To extend the PCP-DV performance evaluation from the single link case, so far considered, to a multi-link network scenario we have considered the network in Figure 7, loaded by multi-hops calls, crossing all the routers, as well as by single hop calls, each loading one link only. We have simulated an homogeneous scenario in which all links have the same capacity, equal to 2 Mb/s. Traffic is generated as in the single link case, with average connection duration equal to 10 minutes. Figure 8 shows the accepted versus offered load and the 99th percentiles of the delay distribution for the two different types of calls and several network sizes (i.e. number of routers). The number l of crossed routers has a negligible effect on the total accepted load, but, as expected, as l increases, we observe a higher percentage of admitted short calls. Indeed, longer calls are more likely to detect high instantaneous loads (corresponding to high rejection probability) in one of the many crossed links. This is an expected behavior of any acceptance control scheme. However, PCP-DV tends to experience a performance degradation for multi-hop calls even in low load conditions, since rejection is also possible in low load conditions, and jitter variance increases with the number of hops. The 99-th delay percentiles, also reported in Figure 8, confirm that, though the multi-hop delay increases with l, the target of a few tens of ms for a backbone can be met.

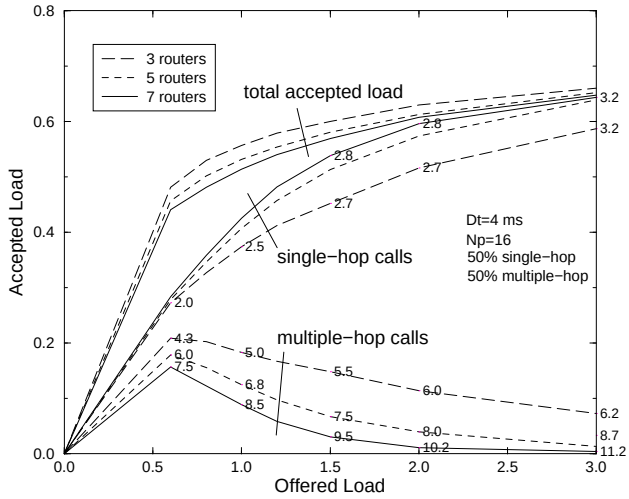

Figure 8: Multi-link scenario: Accepted vs. Offered load for single hop and multi-hop calls and 99th delay percentiles for selected samples; 2 Mb/s links capacity.

5. DELAY VARIATION VERSUS PACKET LOSS

In the previous section, we have shown that PCP-DV is able to guarantee toll-quality performance by pure end-to-end measurements. We now compare PCP-DV performance with those of alternative "pure" EAC schemes, proposed in the literature, which base call acceptance/rejection on the detection of probing packet losses.

To this purpose, we have considered the scheme introduced in [11, 17] and hereafter referred to as SPB (Short Probing Buffer). As in PCP-DV, in SPB N_p probing packets are transmitted during the call setup phase and are forwarded at core routers with lower priority than data packets. However, a very short probing buffer, as short as just 1 or 2 packets, is enforced and probing packets are discarded when the probing buffer is full. The endpoint test rejects a call when the number of lost packets exceeds a predetermined threshold. In our simulations, calls have been rejected whenever one or more probing packets were lost, which is the strictest control that can be exerted on the accepted load, and probing packets have been transmitted at the same rate adopted for PCP-DV. Similarly to PCP-DV, two parameters, N_p and the buffer size buf, regulate the SPB admission mechanism behavior. However, differently from PCP-DV, the buffer size is not tunable by the end points but has to be set in the core routers.

Figure 9 shows SPB's accepted versus offered load for different values of N_p and buffer sizes (buf) ranging between 1 and 2 packets, in a 2 Mb/s single link scenario. The 99th delay percentiles at offered load 4 are also reported. The large spread in the accepted load and the 99th delay percentile underline that the parameter setting in SPB is quite critical. The 0.75 target accepted load can only be attained for buffer size equal to 1 and $N_p = 38$ or buffer size equal to 2 and $N_p = 76$. With larger buffer sizes the accepted load is higher and the QoS achieved is poorer (the corresponding

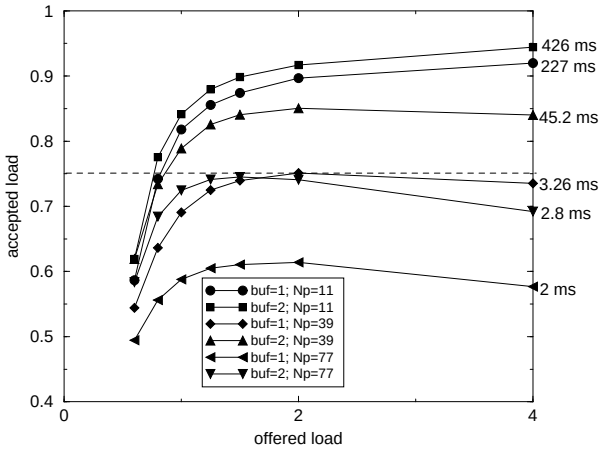

Figure 9: SPB: Accepted vs. offered load for different buffer sizes and number of probing packets. Link capacity equal to 2 Mb/s, average call duration equal to 3 min. 99th delay percentiles at offered load equal to 4 are also reported.

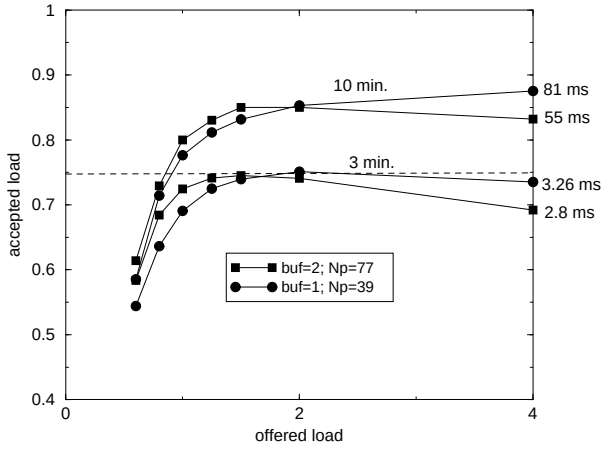

Figure 10: SPB: Accepted vs. offered load for 3 and 10 min. call duration for the cases $N_p = 38$, buf=1, and $N_p = 76$, buf=2.

curves are not reported). It is quite intuitive that managing buffers of such small size is very critical in real routers. Furthermore, Figure 9 shows that, for the curves below the 0.75 accepted load threshold, the accepted load reduces as the offered load increases, suggesting a possible unstable behavior. The reason for this behavior is that in SPB the acceptance/rejection probability is affected not only by the data traffic load but also by the probing load. Indeed, as the offered load increases the probing load also increases, and this in turn translates into probing packet losses (for short buffer sizes) due to the contention among different probing flows.

On the contrary, one should expect that, as the probing load reduces, the SPB rejection probability reduces as well affecting the admission control effectiveness. To verify this effect, we have compared the results obtained with average call duration equal to 3 minutes, with those obtained for 10 minutes calls for which the probing load is reduced by a factor 10/3. In Figure 10 we have compared the results obtained for average call duration equal to 3 minutes, and the two parameter settings ($buf = 1$ and $N_p = 38$, and $buf = 2$ and $N_p = 76$) able to guarantee the 0.75 target load, with those obtained in the 10 minutes case. The results show that as probing traffic decreases, traffic control becomes less effective, and the accepted load significantly increases. This confirms that, unlike PCP-DV, SPB performance is heavily affected by probing traffic load.

Finally, to assess the robustness of "pure" EAC schemes with respect to the adopted traffic model we have slightly modified the Brady's model increasing the average ON and OFF periods to 10 s and 13.5 s, respectively. In these conditions, PCP-DV (see Figure 11) is still capable to provide the same performance as in Figure 6, by just slightly decreasing the jitter threshold D_t. These results, even if obtained in a simple scenario with ON and OFF periods longer than in the voice case, suggest that PCP-DV is able to control also traffic other than voice. It is worth noting that now the probing phase is much shorter than the time dynamics of the traffic model, but the access control scheme is still able to limit the accepted traffic and to guarantee QoS.

On the contrary, SPB is very sensitive to the traffic model. Figure 12 shows that probing phase durations up to 2 seconds ($N_p = 76$) do not succeed in controlling the accepted load, failing to provide toll-quality delay performance, even for the strictest probing buffer setting (1 packet).

The results presented in this section prove that, in an IP telephony scenario, the CAC methods based on delay variations are more effective and tunable than those based on packet losses. Even if not tested in our simulations, we expect that the effectiveness of our scheme improves when dealing with higher rate traffic such as video flows due to the increased number of probing packets. Note that, for the same reason the setting of thresholds in the packet loss based CAC becomes less critical and in this conditions the performance improves [11, 17].

6. CONCLUSION

In this paper we have described the "Phantom Circuit Protocol with Delay Variation" (PCP-DV), a fully distributed end-to-end measurement based connection admission control mechanism able to support per flow QoS guarantees in IP networks. This scheme determines whether a new connection request can be accepted based on delay variations measurements taken on the probing packet at the edge nodes. The PCP-DV approach conforms to a stateless Internet architecture, and it is fully compatible with the Internet architecture promoted by the Differentiated Services framework. The only capability requested to core routers is to implement a 2-priority classes forwarding procedure.

The performance evaluation has shown that tight QoS requirements can be supported by suitably engineering the protocol parameters. We have considered the extremely challenging IP telephony scenario and measured that QoS requirements as tight as just a few milliseconds 99-percentile delay can be guaranteed. The robustness of this mechanism

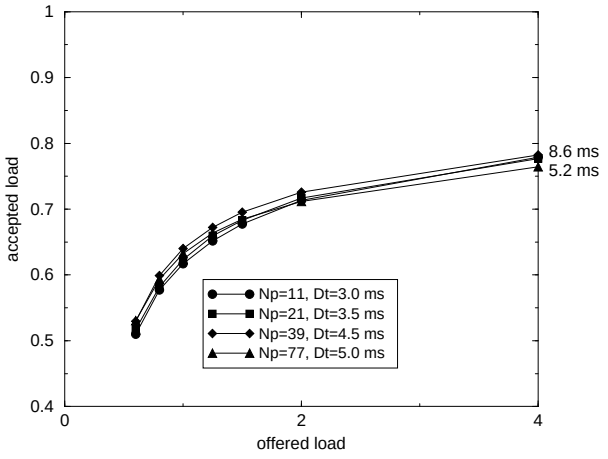

Figure 11: PCP-DV: Accepted vs. offered load for several N_p in a 2 Mb/s link capacity and extended ON-OFF Brady model periods.

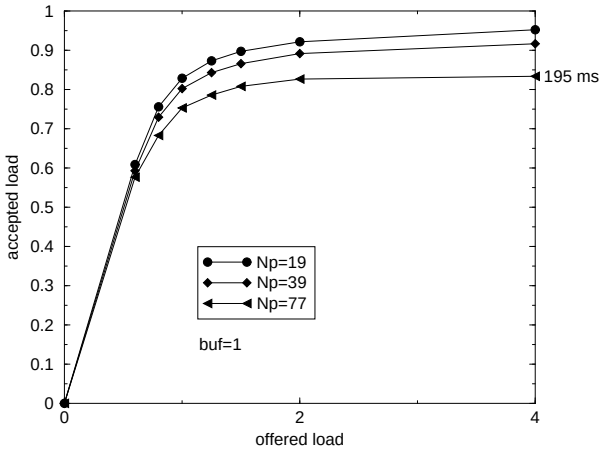

Figure 12: SPB: Accepted vs. offered load for several number of probing packets and buffer size equal to 1 in a 2 Mb/s link capacity and extended ON-OFF Brady model periods.

has also been proven by measuring the performance under several operation conditions.

The comparison with other "pure" end-to-end approaches has proved the effectiveness of adopting the probing packet delay variation, as opposed to packet loss, to measure the network congestion.

7. REFERENCES

[1] P. Goyal, A.Greenberg, C. Kalmanek, W. Marshall, P. Mishra, D. Nortz, K. Ramakrishnan, "Integration of Call Signaling and Resource Management for IP Telephony", IEEE Networks 13(3), pp. 24-32, June 1999.

[2] R. Braden, L. Zhang, S. Berson, S. Herzog, S. Jamin, "Resource Re SerVation Protocol (RSVP)–Version 1 Functional Specification", RFC2205, September 1997.

[3] J. Wroclawsky, "The use of RSVP with IETF Integrated Services", RFC2210, September 1997.

[4] S. Jamin, P. Danzig, S. Shenker, L. Zhang, "A Measurement-Based Admission Control algorithm for Integrated Services Packet Networks", IEEE/ACM Transactions on Networking, 5(1):56-70. Feb. 1997.

[5] W. Almesberger, T. Ferrari, J. Le Boudec, "Scalable Resource Reservation for the Internet", IEEE IWQOS'98, Napa, CA, USA.

[6] M.Grossglauser, D. Tse, "A Framework for Robust Measurement-Based Admission Control", IEEE Trans. on Networking, vol. 7, no. 3, June 1999, pp. 293-309.

[7] L. Breslau, S. Jamin, S. Shenken, "Comments on the Performance of Measurement-Based Admission Control Algorithms", Proc. of IEEE INFOCOM 2000, Israel, March 2000.

[8] K.Nichols, S. Blake, F. Baker, D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and Ipv6 Headers", RFC2474, December 1998.

[9] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss, "An Architecture for Differentiated Services", RFC2475, December 1998.

[10] F. Borgonovo, A. Capone, L. Fratta, M. Marchese, C. Petrioli ,"End-to-end QoS provisioning mechanism for Differentiated Services", Internet Draft, July 1998.

[11] G. Karlsson, "Providing Quality for Internet Video Services", CNIT/IEEE 10th International Tyrrhenian Workshop on Digital Communications, Ischia, Italy, September 1998.

[12] F. Borgonovo, A. Capone, L. Fratta, M. Marchese, C. Petrioli, "PCP: A Bandwidth Guaranteed Transport Service for IP networks ", IEEE ICC'99, Vancouver, Canada, June 1999.

[13] F. Borgonovo, A. Capone, L. Fratta, C. Petrioli, "VBR bandwidth guaranteed services over DiffServ Networks", IEEE RTAS Workshop, Vancouver, Canada, June 1999.

[14] R. J. Gibbens, F. P. Kelly, "Distributed Connection Acceptance Control for a Connectionless Network", 16th International Teletraffic Conference, Edimburgh, June 1999.

[15] C. Cetinkaya, E. Knightly, "Egress Admission Control" Proc. of IEEE INFOCOM 2000, Israel, March 2000.

[16] G. Bianchi, A. Capone, C. Petrioli, "Throughput Analysis of End-to-End Measurement-Based Admission Control in IP", Proc. of IEEE INFOCOM 2000, Israel, March 2000.

[17] V. Elek, G. Karlsson, R. Ronngren "Admission Control Based on End-to-end Measurements", Proc. of IEEE INFOCOM 2000, Israel, March 2000.

[18] G. Bianchi, A. Capone, C. Petrioli, "Packet management techniques for measurement based end-to-end admission control in IP networks", KICS/IEEE Journal of Communications and Networks (special issue on QoS in Internet), July 2000.

[19] F. Kelly, P. Key, S. Zachary, "Distributed Admission Control", Journal of Selected Areas in Communications, December 2000.

[20] L. Breslau, E. Knightly, S. Shenker, I. Stoica, H. Zhang, "Endpoint Admission Control: Architectural Issues and Performance," in Proceedings of ACM SIGCOMM 2000, Stockholm, Sweden, August 2000

[21] Brady, P.T., "A Model for Generating On-Off Speach Patterns in Two-Way Conversation", The Bell System Technical Journal, September 1969, pp.2445-2471.