Admission Control Based on End-to-End Measurements

Viktória Elek, Gunnar Karlsson and Robert Rönngren
Department of Teleinformatics,
Royal Institute of Technology (KTH),
Electrum 204, S-164 40 Kista, Sweden,
{viktoria, gk, robertr}@it.kth.se

Abstract— **This paper proposes a controlled-load service that provides a network state with bounded and well known worst-case behavior. The service is primarily developed for real-time applications. The full system for achieving quality of service to the application consists of an admission control combined with forward-error correction. The admission control is used to limit the packet-loss probability to a known value; the error-control coding (i.e., FEC) is then used to raise the quality above the level enforced by the admission control. The basic idea for the admission control is that a host must probe the path to the receiver before sending actual data. It accepts the session if the probe is received with no or at most a moderate amount of loss. The performance evaluation shows clearly that the proposed scheme avoids network congestion and high packet losses even over short time scales.**

Keywords—**Controlled-load service, measurement-based admission control, differentiated services, integrated services, quality of service, Internet.**

I. INTRODUCTION

Many users' wish lists for improvements of the Internet today might include an item about support for real-time services. The reason is that there exist communication contexts for which the service quality must be higher and more predictable than what is possible to achieve with the available best-effort service. It is clear that users only can tolerate limited amounts of delay and information loss for interactive applications. A communication session is therefore useless if the quality is worse than mandated by those limits. A well-studied example is ordinary voice conversations for which a one way delay of 150 ms is usually not noticeable while it is annoying at 400 ms (ITU-T Rec. G.114). The values are given by human perception and, hence, it makes no difference if the telephony call is carried by a traditional circuit-switched network or by an IP-network. There are similarly limits on the tolerable levels of information loss in the network for voice and video communications. These limits are alas less well defined due to difficulties in judging the impact at the receiver.

Many audio-visual applications are developed to handle the limited amounts of packet loss and delay variation incurred in a lightly loaded network without overly affecting the perceptual quality for the receiver. This means that there ought to be a traffic control in the network that ensures a reasonably bounded load; the adaptation might otherwise be insufficient for providing desired quality to the user. This paper proposes a controlled-load service (CLS) for this purpose along the lines of IETF RFC 2211 [20]. The main idea of the controlled-load is to introduce admission control not available for the best-effort service. The proposed service gives a bound on the packet loss probability, while the delay is limited by the use of small buffers in the network. Forward-error correction is used to recover from packet losses and to raise the loss probability above the level in the network. The service can be implemented without any signaling, and the combination of traffic and error controls allows the acceptable packet loss probability to be selected for each session by the choice of error-control codes.

Corresponding author: Gunnar Karlsson

This paper describes and evaluates the controlled-load service with its admission control based on end-to-end measurements. We do not report any results on the supplementary use of error-control coding for recovery of packet losses. The proposed service requires hosts first to probe the path before starting to transmit data. The connection is admittable if the probe experiences a quality above a specified threshold. The effect of our service definition is that it can readily be included in a differentiated-services architecture [2]. In addition, it has recently been established by Breslau, Jamin and Shenker that many of the hitherto proposed measurement-based admission controls perform equally well irrespective of the complexity of the predictions they use [5]. This lends us support for our simple approach based on spot measurements between sender and receiver. The proposed controlled-load service is a part of a reduced service set architecture that we are developing [14].

The paper is organized as follows. In Section 2 we describe the channel model as well as the admission control based on end-to-end measurement. Section 3 contains the evaluation of the proposed load-control scheme; we summarize related works in Section 4 and outline open questions in Section 5. The paper is concluded in Section 6.

II. THE CONTROLLED-LOAD SERVICE

A. The channel model

The field of data transmission inspires our approach to quality provisioning. A physical transmission channel has a characteristic that can be improved by means of equalization, error-control coding and interleaving. Our analogy is to use traffic control to obtain a reasonably well-defined channel at the network level, equivalent to the physical channel of the transmission system after equalization . The "raw" quality of that channel may then be raised by means of channel coding. The advantages with this combined approach are that an appropriate quality level can be chosen for each session and that the traffic control is much simplified owing to the basic quality offering from the network.

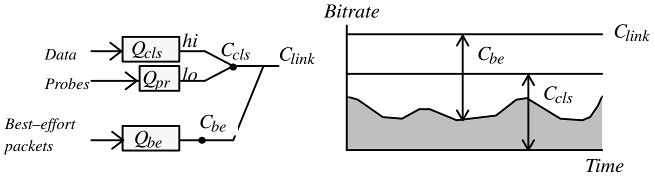

Fig. 1. The service architecture: the controlled-load service is allocated a portion C_{cls} of the link capacity; all remaining as well as allocated but unused capacity is available for best-effort service (white areas). The queue Q_{cls} for data is served at higher priority than Q_{pr}.

The application of error-control codes to a "multiplex channel" offers some problems not encountered in connection with a physical transmission channel. Since the loss is caused by overload, an increase in redundancy has to be compensated by a reduction in source rate to avoid worsening the channel state.

The aim of the admission control is to establish a reliable upper bound on the packet-loss probability in the network. The needed total bitrate as well as the redundancy rate for a session may then be computed given the desired quality at the receiving user, and the maximal packet loss probability.

The actual loss rate may however lie orders of magnitude below the bound. A goal with the traffic control is also to make the variations in loss rate predictable over some reasonable period of time, say a few times the round-trip delay. The loss rate may then be measured by the receiver during the session and fed back to the sender for tuning of the channel code. This allows a session to use the lowest redundancy rate for which the probability of not correctable losses does not exceed a level tolerated by the receiving user.

A small note on the use of error-control codes for packet-loss recovery. Interleaving is required so that the symbols of a codeword get sent in different packets. The packets' sequence numbers indicate the positions of lost symbols in the codewords, which means that they are erasures rather than errors. A given code can correct twice as many erasures as errors. Regular block codes can be tuned in the fashion we need simply by dropping parity-check or information symbols in the codewords. This is called shortening and puncturing, respectively. The decoder has of course to know which parity-check or information symbols have been omitted. It is also possible to use convolutional codes, although they appear to be stronger (and have lower rate) than what is called for here.

B. *The basic principles of the controlled-load service*

The proposed controlled-load service requires its own capacity partition on the links in the network. Unused capacity is fully available to best-effort traffic so there is no waste incurred by the partition (see Fig. 1). The capacity limit for the service cannot be exceeded and may be provided by non-work conservative scheduling [16], which is suitable for real-time transmission with low tolerance to delay variations.

A host is required to probe the network to see if there is capacity available for a new session before sending user data. The probe uses only available capacity within the partition for the controlled-load service class without affecting established sessions. To this end, routers are required to recognize two priorities within the service class. Established sessions are allowed to send packets at the high priority and probes are sent at the low priority. Probe packets are consequently only forwarded on a link if there is capacity left over by the established sessions. The priority is implemented using two queues in parallel served in priority order (as in Fig. 1) or as a threshold based discard of low-priority packets. The performance evaluation assumes the former.

The loss rate that the probe experiences is used to make a decision of whether the session can be established. The probe can be unsuccessful due to temporal network overload or contending probe processes. In this case the session establishment fails, and the sender must back off a random time before issuing a new probe. Persistent overload can be avoided by a proper allocation of capacity to the service class.

Note that only loss and not delay is considered by the admission control for the service; delay is instead limited by the use of small buffers in the nodes which should only provide packet-scale buffering [9].

The description above assumes that the network operator trusts hosts for the probing and the admission decision. If trust cannot be established then the admission control has to be moved within the network proper and the probing has to run between the sender and the exit gateway. The exit gateway returns an encrypted message to the sender, who in turn passes it on to the access gateway where it is decrypted. The access gateway will then take the admission decision and notify the sender of the success or failure of the attempted set-up. With this circuitous message passing, an exit gateway need not know the IP addresses of the access gateways where the probes entered the network. The disadvantage incurred by the lack of trust is the extra functionality needed in access and exit gateways for measuring probe loss, taking admission decision and encrypting/decrypting messages.

C. *The admission control procedure*

Fig. 2 illustrates the phases of a CLS session with acceptance and blocking respectively. First the sender issues a probe: a stream of packets, all of the same length, sent at constant interarrival times. The probe rate should be equal to the maximal bitrate that the sender wishes to use for the session to estimate the available free capacity. The probe rate from a source has to be below a given bound that is settled in a service contract with the operator to limit the sizes of the sessions that can be set up. The maximally allowed rate should be only a small fraction of the capacity allocated to the service class on each link to ensure that statistical multiplexing works well. The packet length is limited in the contract to a value that is reasonable with respect to delay variations.

The sender probes for a time period that it chooses itself. Each

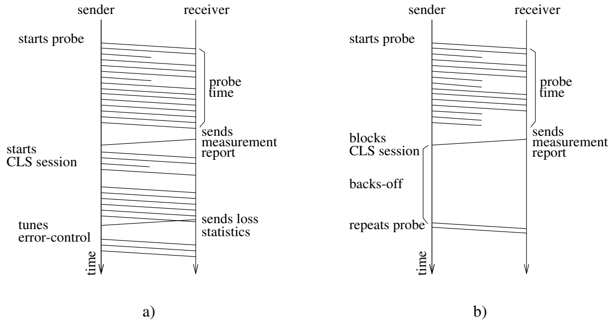

Fig. 2. CLS session with a) session admission and b) blocking.

probe packet includes specification of the probe duration and transmission rate as well as a session identifier. It may also carry information about the session such as specification of the source and channel encodings, but not user data.

The destination counts the received packets until the the probe time period expires. After that it sends a measurement report to the source which consists of the number of probe packets received. The measurement report is carried with high priority to ensure that it is transmitted with low loss.

Based on the measurement report the source decides about the admission. If the calculated probe loss probability is below a threshold, the source may start the CLS session. The threshold is specified in the service contract and all parties sharing the service must use the same value. The receiver continues to measure losses for the established CLS connection and sends periodic feedback to the source for the tuning of the error-control code. The sender may at any time reduce the value of the established bit rate or close the connection. The peak bit rate may only be raised by a new probe to the same receiver.

If the setup is blocked, the source backs off a random time before issuing a new probe. The back-off time is drawn from a uniform distribution of some width, which is doubled for each consecutively blocked attempt to reach the same receiver.

This admission procedure allows a free selection of probe duration. Short probes will however estimate the probe loss probability with large confidence intervals. This means that the probability of acceptance at a given loss level in the network is lower for short probes compared to longer probes. The sender therefore runs a higher risk of blocking by using a short probe but will establish the session faster if successful. Thus, we leave this tradeoff to be settled by the senders for each session setup.

III. The evaluation of the admission control method

The performance evaluation of the proposed admission control addresses the following questions: i) the existence of a general relationship between the probe and session loss probability; ii) the capability of the admission control to avoid congestion; iii) acceptance probabilities of connections with different transmission path lengths.

The network nodes for the controlled-load service are modeled as a system with two priority queues, one for the high priority accepted traffic and the other for the probes. The buffer for the accepted traffic allows packet-scale buffering only. The low priority queue can store a single probe packet and it is only served when the high priority queue is empty.

TABLE I

Parameters and variables used at the performance evaluation.

a:	shape parameter for the Pareto distribution
b_{cls}	buffer size for the controlled-load traffic (7 packets)
b_{pr}	buffer size for the probes (1 packet)
β	location parameter for the Pareto distribution
C:	link capacity (45,150,620 Mbit/s)
d:	packet size (100 bytes)
m:	source mean rate (1.8 Mbit/s)
N:	number of active sessions
p:	source peak rate (5,10 Mbit/s)
p_{cls}:	session packet loss probability
p_{ex}:	measured probe packet loss probability
p_{pr}:	probe packet loss probability
ρ:	utilization factor (session holding time/interarrival time)
s:	number of probe packets transmitted
t:	probe duration (0.5 to 5 seconds)

The parameters and variables used in the performance evaluation are listed in Table I. The presented results are based on mathematical models and simulation, using the sequential YESS simulator [17].

A. The admission-decision rule

A robust admission control should relate the experienced probe loss probability to the expected loss probability of the CLS traffic in a way that is independent of the traffic characteristics. We investigate this relationship for a single multiplexing node.

The traffic for the ongoing sessions is assumed to consist of a number of identical on-off sources with exponentially distributed ON and OFF periods. The probe is represented by a constant bit rate source with a rate equal to the peak rate of the on-off sources.

In the case of packet-scale buffering the packet loss probability due to link overflow can be estimated by the fluid-flow approximation, as:

$$p_{cls} = \frac{E\{(\Lambda_t - C)^+\}}{E\{\Lambda_t\}}, \tag{1}$$

where C is the link capacity, Λ_t the aggregate load, and $(\cdot)^+ = \max(0, \cdot)$ [9]. Using the fluid-flow approximation, the low priority service of the probe means that a probe packet is always dropped in case of link overflow.

Based on Eq. 1 the loss probability for the established sessions p_{cls} and the loss probability of the probe stream, served with low priority p_{pr} are:

$$p_{cls} = \frac{\sum_{i=\lceil C/p \rceil}^{N} (ip - C) \left(\begin{array}{c} N \\ i \end{array} \right) \alpha^i (1-\alpha)^{N-i}}{Nm} \quad \text{and}$$

$$p_{pr} = \frac{\sum_{i=\lceil C/p \rceil - 1}^{N} (p - (C - ip)^+)^+ \left(\begin{array}{c} N \\ i \end{array} \right) \alpha^i (1-\alpha)^{N-i}}{p},$$

where N is the number of established sessions with mean rate m and peak rate p, and $\alpha = m/p$.

Fig. 4 shows the probe packet loss probability as a function of the already established sessions as well as the resulting loss probability if the session would be accepted. The link capacity is 45 Mbit/s, the source mean rate is 1.8 Mbit/s and we consider two burstiness levels with source peak rates 5 and 10 Mbit/s. Simulation results validate the analytical model. For the simulation the traffic of the accepted sessions is modeled as identical, exponential on-off sources, the packet sizes are 100 bytes and the mean ON period lengths 120 packets. Seven packets of the accepted traffic and one packet of the probe stream can be buffered at the node. The measurement time is 20 seconds, the 95% confidence interval is at least one order of magnitude smaller than the measured loss parameter. The probe loss probability is roughly one order of magnitude higher than the session loss probability for both cases and at all load levels.

In Fig. 5 the packet loss probability of the established sessions is plotted versus the probe loss probability for three different degrees of multiplexing and two burstiness levels. This figure suggests that there is relationship between the probe and session losses which does not strongly depend on the traffic characteristics. This relationship suggests that sessions can be accepted if the probe loss probability is at most one order of magnitude higher than the target loss rate of the controlled-load service: $p_{pr} \leq 10 p_{cls}$.

The value of p_{pr} is estimated by the probing process. The accuracy of the estimation depends on the number of samples (probe packets). The more packets that are sent, the closer will the estimated packet loss probability be to the experienced loss rate as well as to the actual loss probability itself.

Assuming uniform probe loss distribution, the probe loss is not higher than the allowed p_{pr}, if the measured loss probability p_{ex} is such that:

$$p_{ex} + z_R \sqrt{\frac{p_{ex}(1-p_{ex})}{s}} \leq p_{pr} \quad \text{and} \quad p_{pr} s > 10, \qquad (2)$$

where s is the number of samples (packets sent), R is the required confidence level, and z_R is the $1-(1-R/2)$-quantile of a normal distribution [15]. The second constraint ensures that the experienced loss is approximately normally distributed.

Fig. 6 shows the p_{ex}/p_{pr} values for different probe durations, for $p_{pr} = (10^{-4}, 10^{-1})$ and for 95% confidence level. The traffic parameters of the probe are p=5 Mbit/s and packet size of d=100 bytes. The acceptable p_{ex} and thus the chance of being accepted increases strongly with the probe duration up to a knee value at $p_{ex}/p_{pr} \approx 0.9$. The sender has however the freedom to use short probing periods but runs a higher risk of being blocked.

B. Congestion avoidance

The aim of the admission control is to avoid high amounts of packet loss due to congestion on the links. We evaluate the proposed solution's capability for congestion avoidance for a single multiplexing node.

The admission control is evaluated at different session blocking probabilities. A session blocking probability in the range of 1 to 5% is considered as the expected operating region for the network during busy hours. We also consider extreme, high overload scenarios that stress the admission control result in a blocking probability of 30%.

The link capacity is 45 Mbit/s and the target maximum session loss rate is 10^{-3} for all the considered cases. The node is loaded by sessions with exponentially distributed interarrival and holding times. The mean session holding time is 2 minutes, the mean interarrival time is set to achieve utilization factors ρ (mean holding time/mean interarrival time) of 10 and 20 that results in the blocking probabilities specified above.

The traffic of the accepted sessions is modeled as identical on-off sources. Exponential and Pareto distributions are considered for the duration of the ON periods, and an exponential distribution for the OFF periods. Packet sizes are 100 bytes for all sessions. The mean rate of the accepted sessions is 1.8 Mbit/s, with peak rate of 5 Mbit/s. The mean length of the ON periods is 120 or 1200 packets, i.e, 20 ms and 200 ms respectively.

Probe durations from 0.5 to 5 seconds are considered. The highest acceptable experienced probe packet loss probability is defined according to the target session loss rate 10^{-3} and the probe duration.

Experiments with a given parameter set were repeated 5 to 10 times, and showed similar features. Results are summarized on time histograms with loss probabilities calculated for 10 second intervals.

The first set of the experiments focuses on scenarios with exponential on-off sources. The mean length of the ON periods is 120 packets and the duration of the probe is 1 second. Fig. 7 shows a 3000 second trace with $\rho = 10$. The blocking probability is 3.37%. Fig. 8 shows a 1000 second trace of the the extreme case with $\rho = 20$. The blocking probability in this case is 34%. The loss probability of the accepted CLS sessions is shown with solid line and of the probes with dotted line. Session blocking events are marked with dashes on the time axis. The target maximum for the packet loss rate of the sessions is shown as well. As the figures show, a loss higher than the target 0.001 is avoided in both cases. Experiments on the overloaded scenario and with 0.5 second and 2 second probe durations give blocking probabilities 38% and 32% respectively.

When the mean length of the ON periods is increased to 1200 packets (Figs 9 and 10), session loss probability higher than

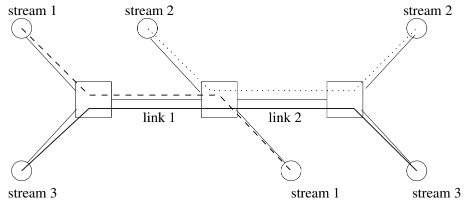

Fig. 3. A three node network to evaluate the effect of long transmission paths.

TABLE II
BLOCKING PROBABILITIES IN THE THREE LINK NETWORK

utilization of stream 1	blocking probability (%)		
	stream 1	stream 2	stream 3
3.33	0.5	11.2	11.8
5	4.7	10.5	18
6.66	9.1	8.7	24.1

0.001 is not avoided in 3 to 5% of the time, and 5 second long probes are necessary to decrease it down to 1 to 2%. Fig. 11 explains why short probes are not accurate in this case. The trace shows a 100 second interval, loss probabilities are calculated for 1 second periods. The probe duration is 2 seconds in this case. Long ON periods lead to bursty aggregate traffic, where the loss probability can change from zero to 10^{-3} in a second. The uniform loss distribution assumed for the confidence interval at the calculation of the acceptable experienced loss probability does not hold in this case, and probes have to sample the loss process over longer time periods.

Figs 12 and 13 show traces when the duration of the ON periods is Pareto distributed according to

$$F(x) = Pr\{X \leq x\} = 1 - (a/x)^{\beta}, \quad a, \beta > 0, x \geq a,$$

where a is the location parameter and β the shape parameter [15]. The shape parameter is $\beta = 1.2$ to have finite mean and infinite variance. The location parameter a is set to get a mean ON period length of 120 packets. 3000 second traces of 10^4 second measurements are shown, to capture the effects of potential long-range dependence. The proposed admission control avoids high maximum losses successfully at $\rho = 10$. Under high utilization the admission control can avoid high losses in 85% of the time, maybe an acceptable result for such extreme cases.

C. The effect of different transmission path lengths

The proposed admission control, as all measurement based solutions, may introduce unfairness by higher blocking of sessions with transmission paths crossing more network nodes. We consider a simple three node network (Fig. 3) to analyze how the path length, or in particular, the packet loss on the transmission path before the bottleneck link affects the blocking probability of the session.

The network is loaded by three multiplexed traffic streams. Stream 1 uses the first and stream 2 the second link only. Stream 3 flows through both links. The utilization factor of streams 2 and 3 is 6.66, generating a bottleneck on the second link. The utilization factor of stream 1 is tuned to values 3.33, 5 and 6.66. Table II gives the blocking probabilities of the three streams under the different loads on the first link. The first case represents a scenario when stream 3 has no remarkable packet loss before the bottleneck link, and the blocking probabilities of sessions in streams 2 and 3 are similar. In the second case there is a light packet loss before the bottleneck link. The effects of the two links add up and the blocking probability of stream 3 is the sum of the blocking probability of streams 1 and 2. In the third case the utilization factors on the two links are equal. As for sessions of stream 3 the losses on the two links sum up, these sessions may be blocked even if packet losses on a single link only do not lead to blocking. As a consequence, the session blocking probability of the two other streams decreases. This third scenario introduces unfairness not present in hop-by-hop measurement based solutions.

To summarize, efficient admission control is possible based only on the peak rate of the new session and the probing of the end to end transmission path. The evaluation of the dynamic behavior of the proposed scheme shows the feasibility of the admission control to avoid network congestion and high packet losses even over short time periods. The results indicate that a needed assumption for the solution is a high degree of multiplexing that reduces load fluctuations on the links. This ensures that the network load does not to change significantly during the probe for the established session.

IV. RELATED WORK - DISCUSSION

Our procedure for the admission control was first proposed by Karlsson in [13]. The work is inspired by the Scalable Reservation Protocol (SPR) developed at EPFL [1]. SPR works in a simple way: a user who wishes to reserve some amount of capacity will send packets with that rate. The packets have a bit set to indicate that a reservation is sought. Each nodes measures the flow of packets with this bit set as well as the flow of packets with reservations. It will forward reservation-making packets at a rate that it is willing to accept and will downgrade the remaining to best-effort service. The receiver measures the rate of reservation packets and notifies the sender of the rate that has been accepted through the network.

There have been proposals similar to ours developed in parallel. For instance, Gibbens and Kelly suggest a marking scheme that can be used for distributed admission control [11]. The main difference with our scheme is that their marking function is more general than the loss function we use for admission decisions. The marking can therefore be used in several modes but it also requires support in the routers.

Borgonovo *et al.* has proposed an admission control for CBR traffic based on end-to-end measurements [3]. The decision is based on measured delay jitter. The group has recently extended the proposal to VBR traffic by suggesting a method for estimation of "peak" available capacity, i.e., the total capacity minus the peak bit rate of the ongoing traffic [4]. The difference with our proposal is that we do not estimate available capacity but rather assess whether a new connection would receive a loss probability below target if accepted.

Measurement-based admission control has been studied by several researchers; some noteworthy contributions are those by Saito [19], Gibbens *et al.* [10] and Jamin *et al.* [12]. The method we propose moves the measurement to the periphery of the network and thus relieves the network nodes of signaling, measurements and admission decisions.

The proposed controlled-load service would fit well within the differentiated-services (diff-serv) framework [2]. Two diff-serv proposals have received much attention: the "allocated-capacity" framework by Clark and Fang [8], and the "two-bit" diff-serv architecture by Zhang, Jacobson and Nichols [21]. Both proposals are interesting and have supported our line of reasoning. However, our scheme is developed for on-demand setup of sessions. This is not supported in the allocated-capacity scheme, and is supported by a new function, a bandwidth broker, in the two-bit scheme.

We are following a new trend of specifying light-weight session set-up protocols. Such protocols are well suited for use in the diff-serv architectures and provide more flexible services than the allocated-capacity and two-bit schemes. Also we base our idea on the work done on measurement-based admission control but have placed the function outside the network for improved scalability.

V. Open issues

There are still many open issues regarding the service specification and our investigation is continuing. The main issues we study are the following.

- The allocation of capacity for the controlled-load service partly determines the blocking probability on the different paths. The background process run by the network management to allocate and de-allocate capacity to the service class is important and its development has not yet been tackled.
- We are going to address the question of mobility and multicast. We believe that the proposed admission control could be modified to work well for scalable multicast sessions and for mobile terminals, since there is no reservation state kept in the network nodes.
- Finally, we would like to implement the scheme and run realistic trials to evaluate the performance further. We are interested in both loss and blocking and it means that we need not only realistic traffic characteristics but also models or traces of the session initiations.

VI. Conclusion

We propose a method to provide service quality for real-time applications by a system that combines tunable forward-error correction and self-admission control. A central aspect of the architecture has been to keep functionality for the load control and quality assurance outside the network. The load control is intended to give a reasonably stable multiplex channel for which the packet loss probability can be estimated end-to-end. This estimate is used to tune the error-control coding to bring the number of uncorrectable packet losses under what is tolerable. The load is regulated by means of an admission control that relies on end-to-end measurements of the network state. The proposed solution provides a flexible controlled-load service by allowing the sessions to select the required service quality through the choice of error-control coding, together with scalability achieved by placing the load control functions at the network boundary. Our performance results show that end-to-end measurements are sufficient basis for admission control.

References

[1] T. Ferrari, W. Almesberger and J.-Y. Le Boudec "SRP: a scalable resource reservation protocol for the Internet", *Computer Communications,* vol. 21, no. 14, pp. 1200-1211, Sept. 1998.

[2] S. Blake, et al., "An architecture for differentiated services," IETF RFC 2475, Dec. 1998.

[3] F. Borgonovo, A. Capone, L. Fratta, M. Marchese and C. Petrioli, "PCP: a bandwidth and delay guaranteed transport service for IP networks," in *Proc. of IEEE ICC'99,*. Vancover, June 1999.

[4] F. Borgonovo, A. Capone, L. Fratta and C. Petrioli, "VBR bandwidth-guaranteed services of Diffserv networks," in *Proc. of IEEE RTAS Workshop,*. Vancover, June 1999.

[5] L. Breslau, S. Jamin and S. Shenker, "Measurement-based admission control: what is the research agenda?," in *Proc. IWQOS*, pp. 3-5, London, UK, May 1999.

[6] R. Braden, D. Clarke, and S. Shenker, "Integrated services in the Internet architecture: an overview," IETF RFC 1633, June 1994.

[7] R. Braden, Ed., "Resource ReSerVation Protocol (RSVP) – Version 1 Functional Specification," IETF RFC 2205, Sept. 1997.

[8] D. D. Clark and W. Fang, "Explicit allocation of best-effort packet delivery service," *IEEE/ACM Trans. Networking,* vol. 6, no. 4, pp. 362-373, Aug. 1998.

[9] "Broadband network teletraffic, Final report of action COST 242," Springer, 1996.

[10] R. J. Gibbens, F. P. Kelly and P. B. Key, "A decision-theoretic approach to call admission control in ATM networks," *IEEE Journal on Selected Areas in Communications,* vol. 13, pp. 1102-1114, Aug. 1995.

[11] R. J. Gibbens and F. P. Kelly, "Distributed connection acceptance control for a connectionless network," in *Proc. of the 16th International Teletraffic Congress*, Edinburgh, June 1999.

[12] S. Jamin, P. B. Danzig, S. J. Shenker, and L. Zhang, "A measurement-based admission control algorithm for integrated services packet networks," *IEEE/ACM Trans. Networking,* vol. 5, no. 1, pp. 56-70, Feb. 1997.

[13] G. Karlsson, "Providing quality for Internet video services,'" in *Proc. CNIT/IEEE 10th Int'l Tyrrhenian Workshop on Dig. Comm.,* pp. 133-146, Ischia, Italy, Sept. 15-18, 1998.

[14] G. Karlsson and F. Orava, "The DIY approach to QoS," in *Proc. IWQOS*, pp. 6-8, London, UK, May 1999.

[15] A. M. Law and W. D. Kelton, "Simulation modeling and analysis," McGraw-Hill, 1991.

[16] M. Mowbray, G. Karlsson, and T. Köhler, "Capacity reservation for multimedia traffics," *Distr. Syst. Eng.,* vol. 5, pp. 12-18, 1998.

[17] R. Rönngren, "YESS/YESP User Manual, Version $\beta 1.2$", Internal Document, Department of Teleinformatics, Royal Institute of Technology, Sweden, March 1996.

[18] S. Shenker, C. Partridge, and R. Guerin, "Specification of guaranteed quality of service," IETF RFC 2212, Sept. 1997.

[19] H. Saito, "Dynamic resource allocation in ATM networks," *IEEE Comm. Mag.,* vol. 35, no. 5, pp. 146-153, May 1997.

[20] J. Wroclawski, "Specification of the controlled-load network element service," IETF RFC 2211, Sept. 1997.

[21] L. Zhang, V. Jacobson, and K. Nichols, "A two-bit differentiated services architecture for the internet," Internet Draft, Internet Engineering Task Force, Dec. 1997. (Available at http://ftp.ee.lbl.gov/papers/dsarch.pdf)

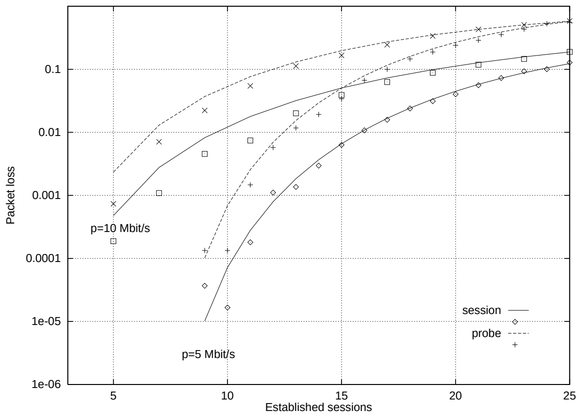

Fig. 4. Session and probe packet loss probability versus number of established sessions. Simulation results are shown with dots. C=45 Mbit/s, m=1.8 Mbit/s, p=5, 10 Mbit/s.

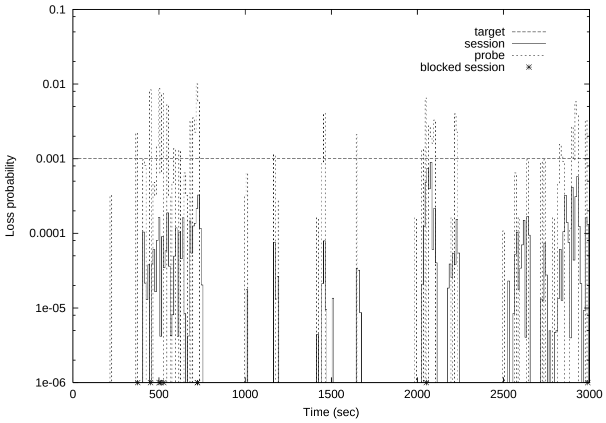

Fig. 7. Session and probe loss probability together with session blocking events. Utilization ratio 10, mean ON period length 120 packets. Measurement time 3000 seconds. (C=45 Mbit/s, m=1.8 Mbit/s, p=5 Mbit/s, $p_{cls} = 10^{-3}$.)

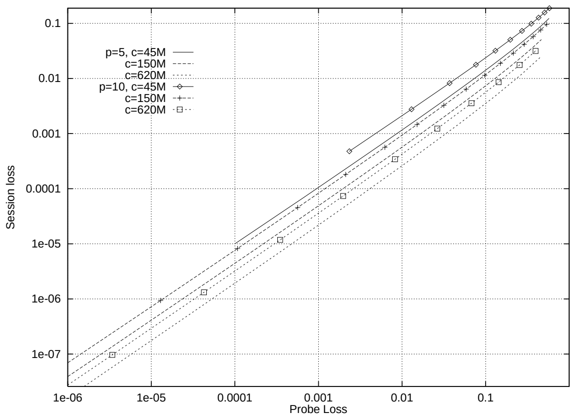

Fig. 5. Session packet loss probability versus probe packet loss probability for different channel capacities and peak rates, with mean rate 1.8 Mbit/s.

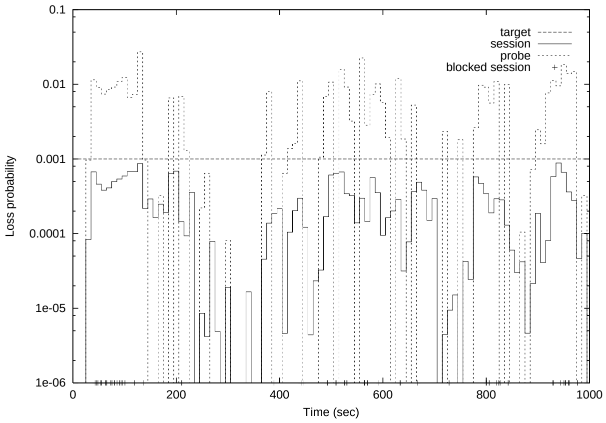

Fig. 8. Session and probe loss probability together with session blocking events. Utilization ratio 20, mean ON period length 120 packets. Measurement time 1000 seconds. (C=45 Mbit/s, m=1.8 Mbit/s, p=5 Mbit/s, $p_{cls} = 10^{-3}$.)

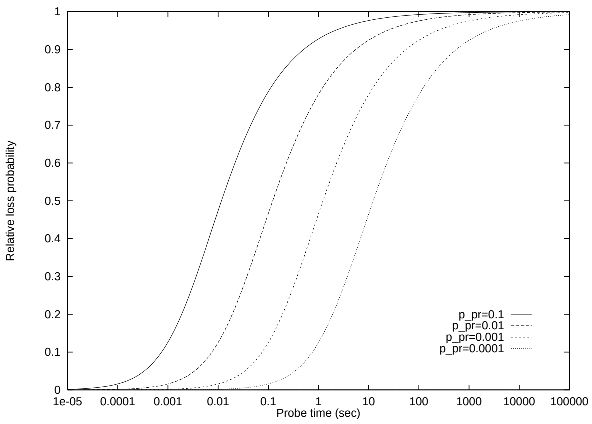

Fig. 6. Relative acceptable experienced loss probability p_{ex}/p_{pr} versus the probe duration in seconds for $p_{pr} = (10^{-4}, 10^{-1})$, p=5 Mbit/s, d=100 bytes.

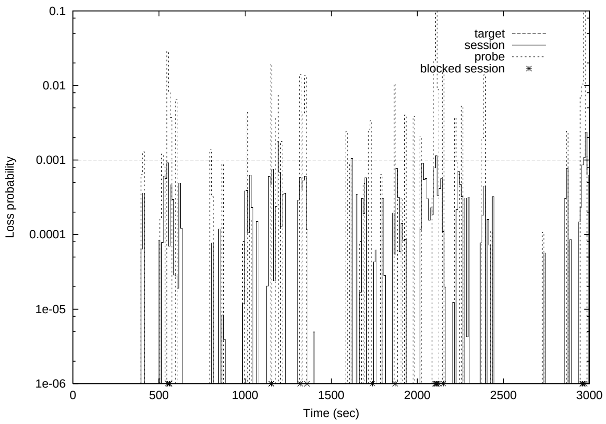

Fig. 9. Session and probe loss probability together with session blocking events. Utilization ratio 10, mean ON period length 1200 packets. Measurement time 3000 seconds. (C=45 Mbit/s, m=1.8 Mbit/s, p=5 Mbit/s, $p_{cls} = 10^{-3}$.)

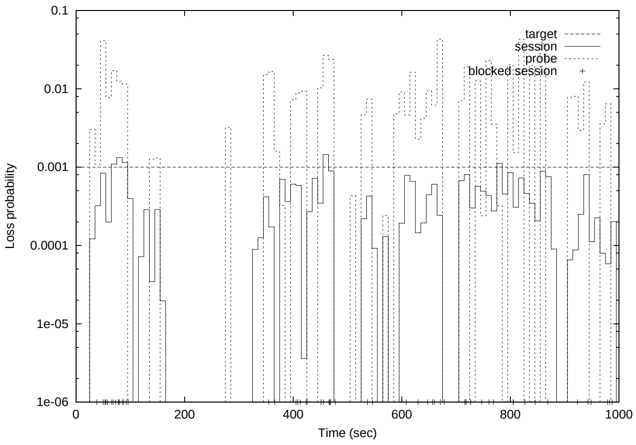

Fig. 10. Session and probe loss probability together with session blocking events. Utilization ratio 20, mean ON period length 1200 packets. Measurement time 1000 seconds. (C=45 Mbit/s, m=1.8 Mbit/s, p=5 Mbit/s, $p_{cls} = 10^{-3}$.)

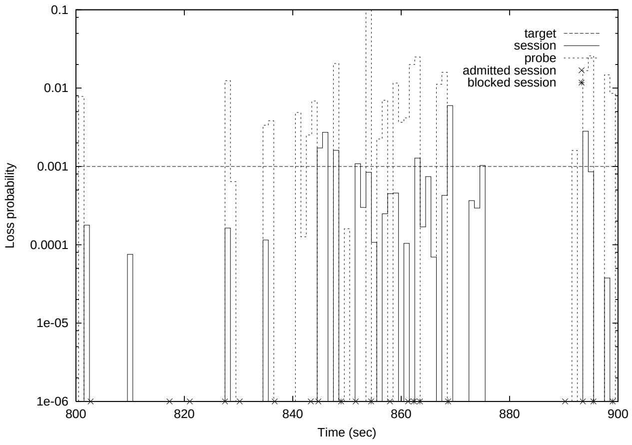

Fig. 11. Session and probe loss probability together with session admission and blocking events. Utilization ratio 20, mean ON period length 1200 packets, probe duration 2 seconds. Measurement time 100 seconds. (C=45 Mbit/s, m=1.8 Mbit/s, p=5 Mbit/s, $p_{cls} = 10^{-3}$.)

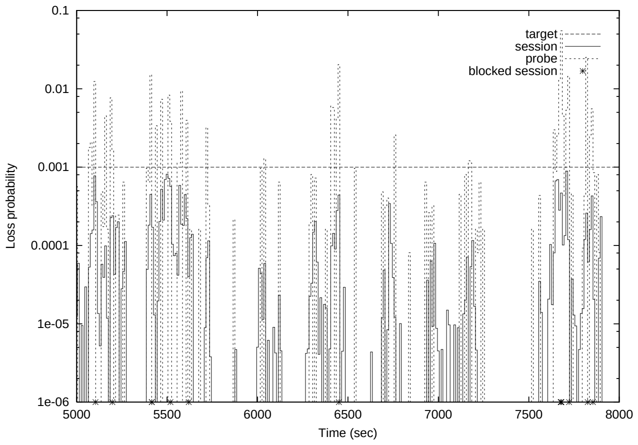

Fig. 12. Session and probe loss probability together with session blocking events. Pareto distributed ON periods. Utilization ratio 10, mean ON period length 120 packets. Measurement time 3000 seconds. (C=45 Mbit/s, m=1.8 Mbit/s, p=5 Mbit/s, $p_{cls} = 10^{-3}$.)

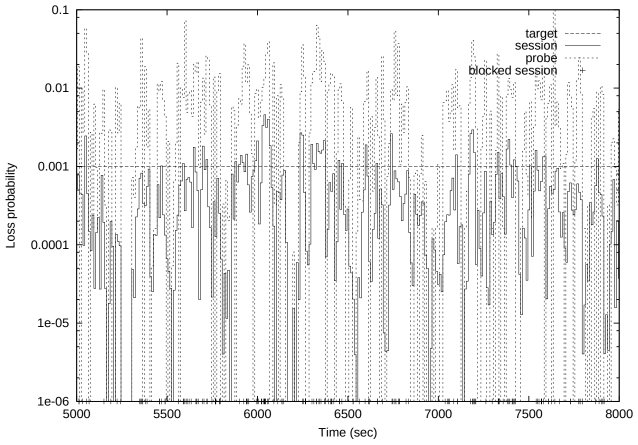

Fig. 13. Session and probe loss probability together with session blocking events. Pareto distributed ON periods. Utilization ratio 20, mean ON period length 120 packets. Measurement time 3000 seconds. (C=45 Mbit/s, m=1.8 Mbit/s, p=5 Mbit/s, $p_{cls} = 10^{-3}$.)