Using Virtual SIP Links to Enable QoS for Signalling

Alexander A. Kist and Richard J. Harris
RMIT University Melbourne
BOX 2476V, Victoria 3001, Australia
Telephone: (+) 61 (3) 9925-5218, Fax: (+) 61 (3) 9925-3748
Email: {kist,richard}@catt.rmit.edu.au

***Abstract*— The Session Initiation Protocol (SIP) will be used on a large scale as a session signalling protocol to provide legacy as well as new services. The 3rd Generation Partnership Project (3GPP) has decided to use the SIP protocol in the IP Multimedia Subsystem of future Universal Mobile Telecommunications System (UMTS) networks. If the SIP protocol is used in carrier grade networks, Quality of Service (QoS) observations are necessary to ensure quality service provisioning. This, in particular, is important since signalling will use multiservice IP transport networks and share its resources with other services.**

This paper proposes the concept of Virtual SIP Links (VSLs) that connect two SIP nodes. VSLs can be used to enable QoS provisioning in SIP signalling overlay networks. Methodologies are introduced to specify, define and dimension these virtual connections. The VSL specification uses the well-known concept of leaky buckets. Simple methodologies are established that are based on known results, to calculate Message Loss Probabilities (MLPs) in leaky buckets. A simple, but efficient, queueing scheme is introduced that reduces the required network resources. Simulation results are given to validate the used models and to underline performance advantages for connections that use VSLs.

I. Introduction

The Session Initiation Protocol (SIP) [1] is a signalling protocol that performs user location, session establishment and other session related tasks in IP networks. It can be used to initiate telephone calls as well as general media sessions. The 3rd Generation Partnership Project (3GPP) [2] has decided to use SIP as the signalling protocol for the IP Multimedia Subsystem (IMS). If the SIP protocol is being used on a large scale in carrier grade networks to provide equivalent telephony services, Quality of Service (QoS) observations are necessary to ensure QoS for customers [3]. These are, in particular important since signalling traffic will share network resources with other network services in multiservice IP networks and the impact of loss or delay of signalling messages is considerably higher than the loss of media packets. Currently there are no signalling specific QoS measures in place.

Transport networks will use QoS methodologies to protect traffic requirements of different services. A combination of Integrated Service (IntServ) [4] technologies at the network edges and Differentiated Service (DiffServ) [5] technologies in the core network appear to provide satisfactory resources [6]. For the remainder of this paper it is assumed that generic methodologies exist and are able to provide QoS resources for signalling. Furthermore, the discussions assume, but do not require that SIP uses the unreliable transport protocol UDP and its own reliability mechanism. The discussions can be adapted for reliable transport protocols like TCP or SCTP if the different protocol timers and resend mechanisms are considered.

Originally, transport services are provided on the basis of conventional IP networks. Any node that is connected to the IP network and has an IP address is globally routable - all nodes are logically full meshed. The same is true for SIP signalling nodes that are connected to the IP network. The IP resources of general-purpose transport networks are also used by services other than signalling. This original network configuration has no dedicated signalling resources.

The introduction of an integrated QoS concept in this context requires the definition of service levels on the SIP layer. Also, the transport network needs to support QoS technologies to guarantee service levels for the message transport. The first step towards an integrated QoS concept on the SIP layer is the definition of virtual SIP connections (VSLs) to allow resource allocation for SIP signalling messages.

A virtual SIP link connects two SIP nodes and is logically located on the application layer. VSLs are defined by their traffic specification (TSpec), i.e. the mean rate, the peak rate, the minimum policed message size, the burst size and the message loss probability. TSpecs can also be used to inform the QoS transport network about the required resources. If the transport network accepts the traffic characteristics, the TSpec defines the connection between two SIP nodes. VSLs and the SIP nodes form the transport independent virtual SIP overlay network (VSON). All relevant issues of the underlying network can be mapped on this layer, for example, bit errors etc. are directly mapped on the virtual SIP links. The SIP network is reduced to well-defined links and nodes. Known methodologies can be applied and new strategies can be developed for VSONs. VSONs also define signalling environments that enable guaranteed service levels.

The objective of this paper is to introduce methodologies to specify, define and dimension these virtual connections. The VSL specifications use the well-known *Leaky Bucket* (LB) concept. Simple methodologies are established to calculate *Message Loss Probabilities* (MLPs) in leaky buckets. A simple and efficient *Delay Line* (DL) queueing scheme is introduced which can reduce required network resources. Simulation results are given to underline performance advantages for connections that use VSL methods.

The remainder of the paper is organised as follows: The next section focuses on the concept of a virtual SIP link. Section III discusses the message loss probability in leaky buckets,

and therefore for VSLs. Section IV introduces the delay line concept. Simulation results in Section V verify advantages of VSLs. The paper concludes with observations that these methodologies can also be used in generic overlay networks and in generic queueing models.

II. Virtual SIP Links

This section elaborates on the concept of virtual SIP links in detail. It outlines the motivation behind the concept, defines VSLs and discusses their operation in 3GPP IMSs. VSL dimensioning and deployment are also discussed.

A. Motivation

The use of virtual SIP links is motivated by three major reasons: To define the resources that are required, to enable sufficiently accurate traffic calculations and to enhance the performance of virtual SIP signalling networks.

To introduce the quality aspect to VSONs, it is necessary to define comparable and predictable parameters that can capture the signalling traffic situation. The mean rate is not enough information since signalling traffic is known to be bursty. More information is required to more completely specify flows in this context. One way of classifying flows is the use of leaky bucket systems. VSLs use LBs and LBs are discussed in Section III

Models, to calculate SIP signalling flows require the input of message loss probabilities, for example, the model in [7] considers only the loss resulting from bit errors. This model can be extended by the message drop probability due to the limited queue size. The VSL concept allows such MLP boundaries. The bursty nature of signalling traffic causes message loss in finite buffers of message transmission queues. Lost SIP messages, on a sustainable level, cause delays due to the timers which are used to detect the loss. Message drops translate into additional delays.

To implement a session level admission control, traffic has to be policed. In this case, the loss is limited to the network edge. The LB, that is part of the VSL definition, can also be used for this purpose. In the case where the message loss is limited to the network edge, the source that sent the original message can use short timers for the first hop. These timers can be below the minimum SIP timer specification of $500ms$. Another alternative could be the use of a message that instructs the source to wait for an arbitrary time before a message is resent. SIP provides this possibility, but the standard allows only clusters of seconds for this purpose. These are too long for this case and, furthermore, it would result in additional messages being sent. Section IV introduces a new way to improve performance that uses the situation where the message loss is limited to the network edge.

B. Definition

VSLs are specified by agreed traffic parameters. These are the mean rate r, peak rate p, the average message size m, the maximum message size M, the bucket depth b, the message loss probability and delay. The *mean rate* reflects the picture of a fluid flow model. It classifies a flow on the basis of bytes per second for a long term average. Where the mean rate describes the average rate, the *peak rate* is, at any given moment, the maximum allowed rate. The normalised arrival rate that is used for the MLP computation, can be calculated by $\lambda_0 = \frac{r}{p}$.

SIP messages will vary in size, so the *average message size* defines the long term average. The longest allowed message size is determined by the *maximum message size*. The *bucket depth* specifies the maximum allowed burst size. The delay variation, due to queueing is determined by this parameter. Messages on the transport network may be subjected to *bit errors* on the transmission media. The message loss probability is the addition of message loss caused by the BER and the message drops due to VSL drops. The next sections explain the operation of VSLs and show methodologies to calculate the parameters that define VSLs.

C. Operation

This section discusses how VSLs operate. The 3GPP IMS is used as an example. Call flows in 3GPP networks require that messages traverse several intermediate SIP proxy servers called *Call Session Control Function* (CSCFs). User clients are connected to P-CSCFs. VSLs connect these P-CSCFs to other inbound SIP servers in the network. In 3GPP jargon, these are I-CSCFs and S-CSCFs. Every server processes incoming messages and routes messages on appropriate VSLs. The number of messages routed on one particular VSL per time unit yields the message arrival rate λ. Before these messages are sent, they are policed by LBs of these VSL. If a message is out-of-profile, i.e. the LB buffer counter exceeds its defined size, the message is dropped. This ensures that the admitted signalling traffic is below the policed peak rate.

If the whole network uses appropriately dimensioned VSLs, message drops are limited to VSON edges. On the basis of available information and by using the methodologies that are introduced in this paper, it is possible to dimension the VSL, so that the number of dropped messages is below a chosen threshold. Note that it is important that the policer drops the out-of-profile packets. Only in this case it is possible to minimize the overall delay which is caused by dropped messages. The next section discusses VSL dimensioning.

D. Requirements Specification

The major VSL dimensioning objective includes the lost messages are below a certain threshold. Constraints include the session arrival rate and other bounding parameters. A VSL is dimensioned for a maximum number of messages per second λ_{max}, allowed on this virtual connection.

The message loss probability depends on the bit error of the communication connection and the MLP that has to be chosen under the consideration of consumer QoS parameters. With the methodologies in [7] the flow size, and therefore the mean rate r can be calculated. For a chosen bucket size b and an average message size of m this yields a normal arrival rate λ_0. Using the equations that are introduced later in this paper a table can be built that maps b and m to λ_0. Once the

normalised arrival rate is known, the required peak rate p can be calculated using Equation (1).

$$p = \frac{r}{\lambda_0} \tag{1}$$

All defining VSL parameters are now available. To finally setup or virtually install a VSL, the traffic specifications have to be accepted by the underlying transport network. The transport network can use these parameters to calculate the required bandwidth.

If a packet is sent between two SIP nodes and it complies with the accepted traffic specification, it will receive the appropriate service levels. The utilisation u is the "current usage" parameter. It can be calculated using Equation (2).

$$u = \frac{\lambda_{max}}{\lambda} \tag{2}$$

The utilisation of a VSL indicates what percentage of resources is used. The requested QoS is guaranteed up to a VSL utilisation of 100%. The next section discusses the deployment of VSLs.

E. Deployment

VSLs require no changes to network hardware. Their functionality can be implemented in a VSL software module which is part of the SIP servers. SIP nodes require VSL sub functions for connections to other SIP nodes, if these connections use VSLs. In principle, the selection of connections that use VSLs is arbitrary. Networks can consist of nodes or areas that use VSLs, and therefore have defined QoS conditions; and areas or nodes that use none, or other methodologies to provide QoS for signalling traffic. If end-to-end QoS guarantees are required, all network regions have to deploy VSLs. Furthermore, paper [8] introduces *Dynamic Resource Allocation* (DRA) for VSLs. The next section discusses the models that analyse the dependency between λ_0 and the message loss probability.

III. VSL MLP MODELS

VSLs use the well-known leaky bucket scheme to define their traffic specifications. Originally, LBs have been proposed as a mechanism to control the cell arrival process in ATM systems (e.g.[9]). Later, the IETF adapted the concept to define traffic specifications (TSpec) in the IntServ framework. In principle, all QoS control service technologies can use LB parameters to describe the nature of bursty traffic. These parameters are defined in RFC 2215 [10].

The following terminology is commonly used: The mean rate r is called the average/token rate (IETF) or sustainable cell rate (ATM). The mean rate is a theoretical long-term average inter-arrival time in respect to the link speed and is measured in bytes per second. The peak rate p (IETF/ATM) limits the theoretical minimum inter-arrival time between packets. It describes the limit of the traffic source. It is measured in bytes per second. The bucket depth (IETF) or burst tolerance (ATM) b describes the maximum amount by which the source is allowed to burst at the peak rate. It is measured in bytes. The average message size m is measured in bytes and describes the long-term average of the packet size. The maximum message size defines the maximum number of bytes allowed per packet. The arrival rate λ is measured in arrivals per second.

LBs are well understood and can be used to police peak rates and shape mean rates [11]. This section introduces simple methodologies to calculate the message drop probability in finite size buckets. In the past, much research has focused on this topic. Various models are published that enable the calculation of MLPs for LB with utilisations below one. Here the calculations use two models. The first model describes loss probabilities for leaky buckets with utilisation of one. This is based on concepts that are used by the Erlang formulas. The other model describes the case for utilisations below one and can be found in [12]. Simplifications are proposed for the later model.

Duffy et al. report in [13] that traffic at the message level exhibits long range dependencies, but Skoog notes in [14] that for Poisson call arrivals an M/G/1 model gives a good approximation for SS7 signalling link queues when the utilisation is below unity. The message drop estimations in this paper assume Poisson arrivals since the session arrivals are assumed to be Poisson and the memory of the system is very short - only a few packets. Related messages that belong to the same call flow are separated by the round trip time. RTT is much shorter, than the time messages spend in queues. Note that this assumption only impacts on the MLP, calculation but not on the VSL concept.

A. MLP for a Utilisation of One

This section uses the concept of a *Small Leaky Bucket* (SLB) to define a loss calculation methodology. A SLB consists of a single buffer space and a single server ("bucket hole"). The buffer size ("bucket size") is equal to the maximum allowed packet size M. This bucket is served with a rate of λ_{SLB} (leak rate). The service time for one packet of mean packet size m is given in Equation (3).

$$T_{SLB} = \frac{1}{\mu_{SLB}} = \frac{m}{C} \tag{3}$$

The equivalent capacity of the SLB is C and μ_{SLB} is the service rate of the SLB. The traffic unit of "packet calls" is then $A = \frac{\lambda_{SLB}}{\mu_{SLB}}$. A packet is lost if more than one packet arrives while the current packet is being served. If the model considers the special case where the leak rate λ_{SLB} is equal to the arrival rate λ of a Poisson process, the message loss probability P_{SLB} can be calculated using Equation (4).

$$P_{SLB} = 1 - \left(e^{-\lambda \cdot T_{SLB}} + \lambda \cdot T_{SLB} \cdot e^{-\lambda \cdot T_{SLB}}\right) \tag{4}$$

This equation describes the probability that more than one new packet arrives during the service period. The analogy to Erlang equations can be observed: The first term in the parentheses shows the probability that "no calls arrive"; and the second term, the probability that "one call arrives". In all other cases the messages are dropped. If the leak rate corresponds to the long-term mean rate of the arrival process, the utilisation is one. The normalised arrival rate λ_0 is equal to the utilisation

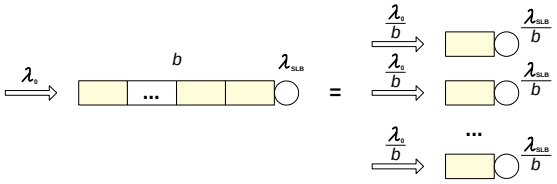

Fig. 1. SLB Model

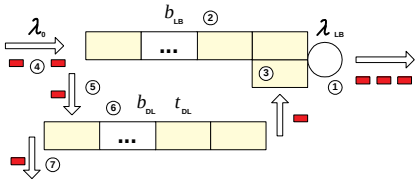

Fig. 2. Delay Line - Operation

of the bucket and is unity, in this case as well. It can be calculated using Equation (5).

$$\lambda_0 = \lambda \cdot T_{SLB} \tag{5}$$

where λ is the arrival rate of the messages at the LB and T_{SLB} is the service rate. Equation (6) depicts a simplified version of Equation (4) using the normalised arrival rate.

$$P_{SLB} = 1 - (1 + \lambda_0) \cdot e^{-\lambda_0} \tag{6}$$

This model can be extended to a LB with a bucket size b in packets, under the assumption that the utilisation remains one. In this case, the loss probability can be calculated by the sum of the loss probability of b SLBs that are served at a rate $\frac{\lambda}{b}$. This is possible because the sum of a Poisson process yields a Poisson process. Figure 1 depicts this model graphically. A bucket of size b served at a rate λ_0 is shown on the left hand side. The packet flow can be split into b equal parts and can be served by b different SLBs at a rate of $\frac{\lambda_0}{b}$. Equation (7) shows the loss probability P_{LB} calculation in this case.

$$P_{LB} = b - (b + \lambda_0) \cdot e^{-\frac{\lambda_0}{b}} \qquad for \quad \lambda_0 = 1 \quad or \quad b = 1 \tag{7}$$

Note that Equation (7) only provides valid results in the case that either the normalised arrival rate (utilisation) $\lambda_0 = 1$ or the buffer size $b = 1$.

B. MLP for Utilisations Below One

The limitations of the previous discussed model was that it required either a λ_0 of one or a buffer size b of one. There are various models available that approximate the drop probabilities of queues under the assumption that λ_0 is smaller than one. These models usually provide poor results for $\lambda_0 = 1$.

The discussion in this section uses the approach described by Pitts and Schormans [12]. They use the definition of an instantaneous excess rate to decide if a packet is served or if it has to be queued. Then, they connect the arrivals of the excess rate packets via balance equations. Their derived result is given in Equation (8).

$$P_a(\lambda_0) = \left(\frac{\lambda_0 \cdot e^{\lambda_0} - e^{\lambda_0} - \lambda_0^2 + \lambda_0 + e^{-\lambda_0}}{\lambda_0 - 1 + e^{-\lambda_0}} \right)^{b+1} \tag{8}$$

It shows the probability $P_a(\lambda_0)$ that the queue exceeds b packets. As above, the arrival rate λ_0 is the arrival rate λ normalised by the service rate of the observed system. The interval of interest [0,1] for λ_0 also yields an interval of [0,1] for the loss probabilities P. Possible simplifications have to be valid in these intervals.

The simplest approximation for the term in parentheses is λ and therefore the linear function $P_d(x) = \lambda_0^{b+1}$. A better approximation is given by: $P_c(x) = (\frac{\lambda_0}{2-\lambda_0})^{b+1}$. The maximum error for $b = 0$ of this function compared with the original function $P_a(\lambda_0)$ is about 4%. The third empirical simplification scales function $P_d(x)$ with e^{λ_0} and normalises it with e to keep the boundaries of 0 and 1. This yields Equation (9).

$$P_b(\lambda_0) = \left(\frac{\lambda_0 \cdot e^{\lambda_0}}{e} \right)^{b+1} \tag{9}$$

The error function $Err(\lambda_0)$ in this case for the basis can be calculated using Equation (10).

$$Err(\lambda_0) = \frac{\lambda_0 \cdot e^{\lambda_0} - e^{\lambda_0} - \lambda_0^2 + \lambda_0 + e^{-\lambda_0}}{\lambda_0 - 1 + e^{-\lambda_0}} - \frac{\lambda_0 \cdot e^{\lambda_0}}{e} \tag{10}$$

The extreme points of the error function show that the simplification error in this case is in the worst case 0.4%. The simplification is therefore adopted as sufficiently accurate. The combined loss for both MLP models can be calculated by the minimum of Equation (7) and Equation (9).

IV. Delay Line

This section introduces the simple Delay Line (DL) methodology. It can reduce the overall delay due to the lost messages and it can be used to increase the utilisation of VSLs. Figure 2 depicts a simplified schema for the concept. It consists of one server (1) and a buffer b_{LB} (2). These parts are identical to conventional leaky buckets. Additionally, it has an extra priority buffer space (3). If this buffer is occupied it is served before the normal buffer (2).

If a packet arrives at a LB (4) and the buffer b_{LB} is full the packet is dropped (5). In this scheme the packets are not dropped, they are delayed by a delay line (6). It is of size b_{DL} and delays the packet by a constant fixed value of t_{DL}. If the number of packets in the delay line exceeds its size, the excess packets are dropped (7). The maximum loss for this setup can be calculated by the equations that were introduced in Section III, if a buffer size of $b = b_{LB} + b_{DL} + 1$ is used. This is based on the assumption that for the worst case all available buffer spaces are occupied, and that this is independent of their location.

The advantage of this scheme compared to simply increasing the buffers is that, in a practical case, the buffer b_{LB} will be located in the network and define the virtual buffer size as part of the traffic specification. It is located in the network and therefore an expensive commodity. On the other hand,

the buffer that is part of the delay line will be located in local nodes at the network edge, where buffer space is cheap.

In this practical case, the scheme is slightly different from Figure 2. At the sender side the buffer b_{LB} is implemented with a counter and packets are not buffered. The site has also no access to the server, therefore, buffer (3) cannot be implemented. In this case, packets that leave the delay line are directly transmitted. Using the VSL concept combined with a DL can significantly reduce random fluctuations of the traffic, and therefore reduces the required utilisation and/or bucket size. Section V gives simulation results that graphically underline the advantages of this scheme.

To use DLs, certain assumptions have to be fulfilled. To avoid messages being received out of sequence, the delay t_{DL} has to be lower than the inter-arrival time between two consecutive packets in the same transaction. Furthermore, if the service in the network is much better than it was specified in the SLA that described the VSL, messages are unnecessarily delayed.

V. VSL Simulations

This section provides simulation results that compare performance parameters of VSLs. A discrete event simulator for SIP networks with 3GPP-like topologies was used to obtain these results. It uses the Mersenne Twister as a random number generator. For a large number of requests (100,000) the absolute round trip delay was measured, i.e. the time between the instant a request is sent and the time instant the response is received. This also included time out and resent messages in the case of losses.

For simplicity, a call flow with only one request and one response was used. Both messages had to pass 7 intermediate proxy servers (This is based on 3GPP IMS setup). The session arrival rate in this simulation followed a Poisson process. The mean arrival rate was set to be 100 sessions per second. The mean message size was uniformly distributed between 300 and 700 bytes. The propagation delay of all intermediate links (18), added up to 800 ms.

Three different cases were simulated: Random message losses in the servers, the use of VSLs and the combined use of VSLs/DLs. The results are depicted in Figure 3, Figure 4 and Figure 5 respectively. These delay histograms use logarithmic scales. The x-axes depict delay bins in ms. The step size is two ms. The y-axes depict the respective counts.

In the first case, a random message loss was simulated. A random loss in this context reflects a message loss that is due to finite size buffers. An overall loss probability of 1.18% of all sessions was split between the servers. For practical cases, this number was too high, but to be able to simulate the network in a reasonable time with reasonable statistical accuracy, this number was chosen. In a realistic situation, this overall loss probability will be much significantly lower.

The first peak in this histogram indicates messages that were not lost. To give an indication of the statistical accuracy of these results, the 95% confidence interval is given. The probability that no messages are lost and therefore, the

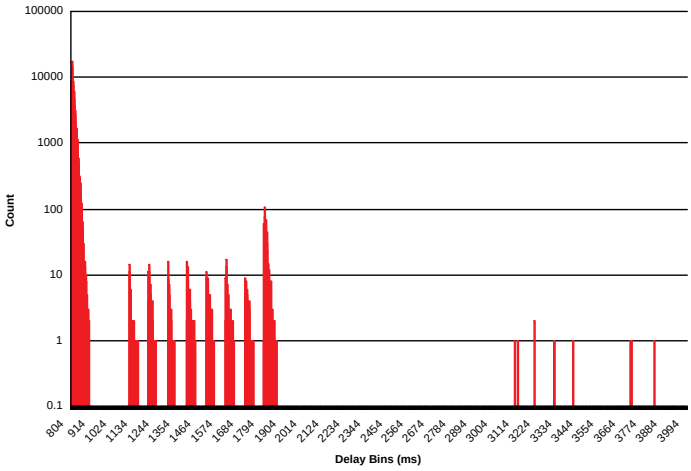

Fig. 3. Delay Histogram - Simulated Random Loss

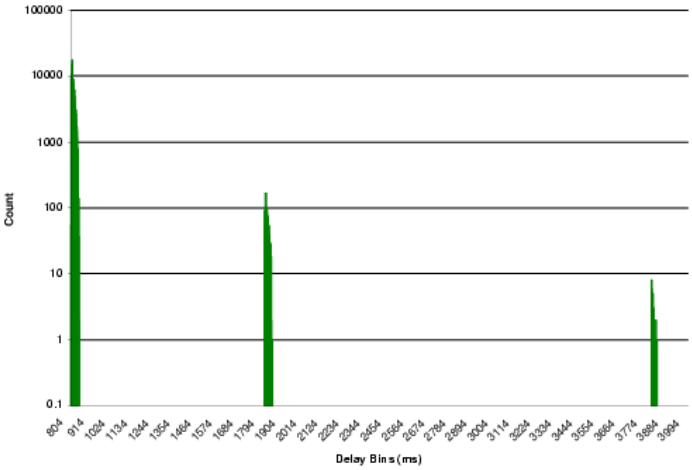

Fig. 4. Delay Histogram - Using VSL

probability that they are located in the first peak yields a range of [98.81%,98.84%].

The next peak indicates messages that were lost on the reverse connection between the first and the second node, whilst the third peak indicates messages that that were lost on the reverse connecting between the third and the second node and so on. The peak at 1900 ms is due to requests that were lost on any of the forward links. Peaks further to the right account for messages that were lost twice or more.

The histogram depicted in Figure 4 was generated using the same principal setup but no random loss occurred and the first link in the connection used the VSL concept. Because the transport network accepted the VSL traffic specifications, the traffic encounters no loss due to queueing in the transport network. The VSL was dimensioned in a way that the loss is the same as for in the first graph. It had a buffer size of 2500 Bytes (5 spaces), a mean rate of 5000 Bytes and capacity of 75000 Bytes. This is equivalent to a normalised arrival rate λ_0 of $66,6\%$. In this case, Equation (9) yields 1.18% loss.

The loss is limited to the VSL itself, and therefore the first hop. The 95% confidence interval in this case is

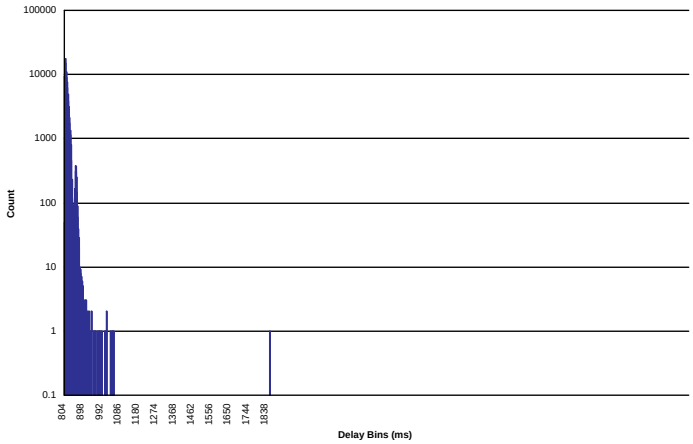

Fig. 5. Delay Histogram - Using VSL and DL

[98.82%,98.88%], which agrees with the analytically calculated loss. The first peak in Figure 4 depicts messages that encountered no loss; the second peak indicates once lost messages and the third peak indicates message that were lost more than three times. By itself, this does not provide a considerable performance improvement. But in this case, it is possible to use much shorter timers for lower layer resend mechanisms.

Another option is that VSLs use DLs. This case is depicted in Figure 5. Again the setup was the same as for the first two cases. Additionally, a DL with 50 ms delay was used. Delays that are due to the resending of lost messages which, in turn, are caused by the limited number of resources, are greatly minimised. Practically all sessions have a delay below 1000 ms. This shows that the use of VSLs with DLs can improve the performance considerably and ensure QoS for signalling traffic. The next section outlines areas of further work concerning these concepts.

VI. Further Work

In three major areas, further work is required. Firstly, the underlying transport network: The possible implementation and interaction with underlying transport technologies requires further attention. This includes impacts on the VSL scheme by QoS provisioning technologies like DiffServ, RSVP etc. Secondly, MLP calculation: If the traffic specifications have costs assigned to each parameter, an optimisation is possible, for example finding the set of parameters, where MLP is below a threshold so that the costs of the required resources are minimal. Lastly, generic overlay networks: The methodologies of VSLs and particularly the DL concept are not specific to SIP signalling networks. Further research can focus on the usability of these concepts in generic networks.

VII. Conclusion

This paper defined the concept of virtual SIP links. It applied the methodology of a leaky bucket to define VSLs on the SIP layer. By using this concept it is possible to define a SIP overlay network that consists of SIP nodes and VSLs. The chapter provided methodologies that enable the dimensioning of the virtual connections and introduced simple methodologies that can predict the drop probabilities in VSLs which use leaky buckets.

Furthermore, the concept of a delay line was introduced which enables the use of additional queueing buffers, located outside of the transport network. The combination of these concepts can improve performance considerably, as delay simulations have shown. The methodologies discussed in this chapter can also be applied to generic emerging overlay network technologies.

One major advantage of VSLs is the possibility of dynamic resource allocation which is addressed in [8]. Since the resources are requested from underlying IP networks it is possible to implement schemes that adapt the traffic subscription to the current requirements.

VIII. Acknowledgements

The authors would like to thank Ericsson AsiaPacificLab Australia and the Australian Telecommunications Cooperative Research Centre (ATcrc) for their financial assistance for this work. The helpful comments by Dr. Bill Lloyd-Smith in the CATT Centre, RMIT University, are gratefully acknowledged.

References

[1] J.Rosenberg, H.Schulzrinne, G.Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler, *SIP: Session Initiation Protocol*, IETF, June 2002, RFC 3261 (Obsoletes: RFC 2543).

[2] 3rd Generation Partnership Project, *About 3GPP*, October 2002, http://www.3gpp.org.

[3] A.A. Kist and R.J. Harris, "SIP signalling delay in 3GPP," *In Sixth International Symposium on Communications Interworking of IFIP - Interworking 2002, Fremantle WA, October 13-16*, 2002.

[4] R. Braden, D. Clark, and S. Shenker, *Integrated Services in the Internet Architecture: an Overview*, IETF, June 1994, RFC 1633.

[5] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, *A framework for differentiated services*, IETF, December 1998, RFC 2475.

[6] L. Zheng, A. Dadej, and S. Gordon, "Hybrid quality of service architecture for wireless/mobile environment," *In Proceedings of Sixth International Symposium on Communications Interworking of IFIP - Interworking 2002, Fremantle WA, October 13-16*, 2002.

[7] A.A. Kist and R.J. Harris, "A simple model for calculating SIP signalling flows in 3GPP networks," *In Second IFIP-TC6 Networking Conference 2002, Pisa, Italy May 19-24*, May 2002.

[8] A.A. Kist and R.J. Harris, "Dynamic resource allocation in 3GPP SIP overlay networks," May 2003, under submission.

[9] E. P. Rathgeb, "Modeling and Performance Comparison of Policing Mechanisms for ATM Network," *IEEE Journal of Selected Areas in Communications*, vol. 9, no. 3, pp. 325–334, April 1991.

[10] S. Shenker and J. Wroclawski, *General Characterization Parameters for Integrated Service Network Elements*, IETF, September 1997, RFC 2215.

[11] V. Anantharam and T. Konstantopoulos, "Burst reduction properties of the leaky bucket flow control scheme in ATM networks," *IEEE Transaction on Communications*, vol. 42, no. 12, pp. 3085–3089, December 1994.

[12] J. M. Pitts and J. A. Schormans, *Introduction to IP and ATM Design and Performance: With Applications Analysis Software*, Wiley, Chichester, 1996.

[13] D.E. Duffy, A.A. McIntosh, M.Rosenstein, and W.Willinger, "Statistical analysis of CCSN/SS7 traffic data from working CCS subnetworks," *IEEE Journal of Selected Areas in Communications*, vol. 12, no. 3, pp. 544–551, April 1994.

[14] R. A. Skoog, "Study of clustered arrival processes and signaling link delays," *In Teletraffic and Datatraffic in a Period of Change (Proc. 13th ITC, Copenhagen, 1991)*, pp. 61–66, 1991.