VoIP Network Dimensioning using Delay and Loss Bounds for Voice and Data Applications

Veselin Rakocevic
School of Engineering and Mathematical Sciences
City University, London, UK
V.Rakocevic@city.ac.uk

Robert Stewart, Ronan Flynn
Athlone Institute of Technology Athlone, Ireland
rstewart@ait.ie; rflynn@ait.ie

Abstract: This paper analyses resource provisioning for enterprise Voice-over-IP (VoIP) networks. Simulation and analytical methods are used to enhance the provisioning process, which ultimately aims to provide the delivery of consistent Quality of Service (QoS). Consistent QoS is increasingly important in order to deliver an adequate service, as defined in terms of Service Level Agreements (SLAs). This paper defines simple guidelines for network dimensioning in a multimedia environment in terms of end-to-end delay for the voice traffic, and in terms of throughput and packet loss for TCP data traffic. A realistic environment is modeled and simulated using ns-2. The model consists of a prioritized network in which intermediate routers perform priority scheduling to provide differentiation of Internet services.

1. Introduction

The increase in both popularity and capacity of the Internet has led to the increasing need to provide real-time voice and video services to the network. While the potential benefits of these services are enormous, the process of adapting the connectionless data-oriented design of IP networks to real-time traffic is rather slow. Numerous technological problems that network designers are faced with are all focused on the issue of Quality of Service (QoS). The main QoS problem in the Internet is how to provide (and guarantee) the bandwidth, delay and packet loss bounds to the real-time network traffic. Other QoS problems include end-to-end QoS deployment, network security, and network reliability. A range of different service classes will be provided in the next-generation Internet, and this has focused attention on providing more detailed and accurate Service Level Agreements (SLAs), including bounds on end-to-end delay and packet loss.

With the increasing focus on traffic prioritization to support voice-data integration in corporate intranets, methods are needed to dimension limits on end-to-end delays. Certain "soft" real-time applications, such as interactive packetized voice and video, are delay sensitive but loss tolerant [6] hence the end-to-end delay distribution seen by packets in a session becomes a critical performance measure. This is particularly important for the provisioning of delay sensitive services such as VoIP applications, typically provided over RTP/UDP/IP. The introduction of VoIP has enabled many organizations to supplement or replace existing circuit-switched telephony networks, providing new value-added services which include Internet call waiting, IP conference calls managed from the desktop, voice access to web content and IP-based 'virtual PBXs' for small businesses.

On the other hand, network provisioning for real-time services needs to be performed with some consideration for the end-to-end performance of data applications, which are typically transmitted over TCP. This paper provides a novel approach to network

dimensioning and QoS analysis by taking into considerations not only requirements of real-time services, but also the requirements and the end-to-end network performance of data applications.

This paper analyses resource provisioning for enterprise VoIP networks. It follows a similar analysis [1] in which guidelines for network dimensioning for small VoIP networks were given. Previous work [1] analyzed enterprise VoIP networks and showed the limitations of using the M/D/1 model for accurate dimensioning of SLAs end-to-end in the VoIP network. Traffic smoothing, the use of large buffers and admission control were identified as traffic control methods which aim to maximize the number of VoIP flows accepted for a given end-to-end delay bound in a VoIP network. The work in [1] defined the recommended buffer utilization for SLA dimensioning in enterprise networks ranging from 10-100 multiplexed VoIP flows.

The work presented in this paper extends the work in [1] by investigating the end-to-end delay performance of VoIP applications, and at the same time the throughput of TCP data applications. We observe both the cases of FIFO scheduling and prioritized scheduling on the basis of the CBQ/WRR model. Section 2 describes the network model in more detail.

We argue that there must be a balance in the decision how much we suppress the data traffic in the Internet in order to provide space for the voice traffic [8]. Naturally, the exact solution for this problem will also include the economics of the system – it is likely that more expensive services will have priority over the cheaper services, and it is likely that real-time services, including voice, will be more expensive than the data services.

2. The Network model

The network model, given in Fig. 1, is general enough to enable us to form network dimensioning conclusions. The observed network model consists of a sequence of 10 network nodes (routers) with finite buffers. The routers have an option of implementing priority scheduling of the incoming traffic. The simulation is performed using the ns-2 simulator [4].

Two basic traffic types exist in the network, the real-time traffic and the elastic data traffic. The real-time traffic (VoIP) uses RTP and UDP, while the data traffic is modeled as file transfer over TCP. VoIP sources are modeled as on-off sources with exponential distribution of on and off times. The parameters for the VoIP and data sources are given in Table 1. The chosen parameters for the VoIP on-off sources are widely used in the literature, while those of the data were chosen to reflect the behaviour of data sources that are bursty with a high peak rate and long mean off durations [9]. The background traffic at each node is modeled as short-lived on-off data transfer over TCP. The TCP traffic is modeled as 'background' traffic - each of the data transfers passes through only one buffer. VoIP sources are transported through all 10 buffers, and their end-to-end performance (delay) is monitored.

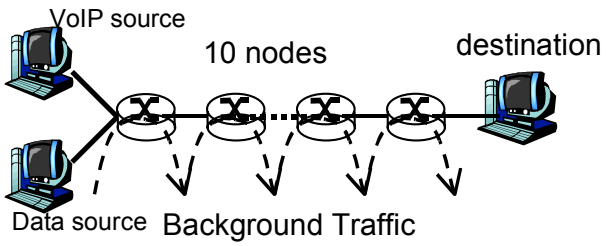

Figure 1 Network Model

In the model QoS is achieved through the use of priority schedulers. The priority schedulers in our model use CBQ/WRR scheduling. Class-based Queueing (CBQ) was introduced by Jacobson [2], and is used to meet the link sharing requirements in the differentiated network environments. The CBQ/WRR scheduler takes the incoming packets from a number of traffic classes and schedules them according to the weights associated to the classes. This introduces resource partitioning [7] in the network model, and raises the issue of optimal allocation of weights in the priority scheduler to achieve maximal network performance. Dimensioning the prioritized network then is closely related to optimal weight allocations.

Voice-	
mean ON duration	0.35 seconds
mean OFF duration	0.65 seconds
Peak rate	167 packets/s
Data-	
mean ON duration	0.5 seconds
mean OFF duration	1.9 seconds
Peak rate	1000 packets/s

Table 1 Simulation parameters for the multiplex of ON/OFF sources

3. Simulation Modeling

This research uses simulation modeling mainly because the scale of the network in Figure 1 makes it analytically intractable end-to-end. In network analysis, it is well known [3] that packet interarrival times become strongly correlated with packet lengths once packets have traveled beyond their entry queue. Alternative methods have been used for analysis of such a network model by the authors in previous work. The accuracy of using the delay distribution from a single node and applying convolution to find the end-to-end delay is given in [5]. Results from [5] compare the theoretical end-to-end delay results with results from simulation and are shown in Figure 2.

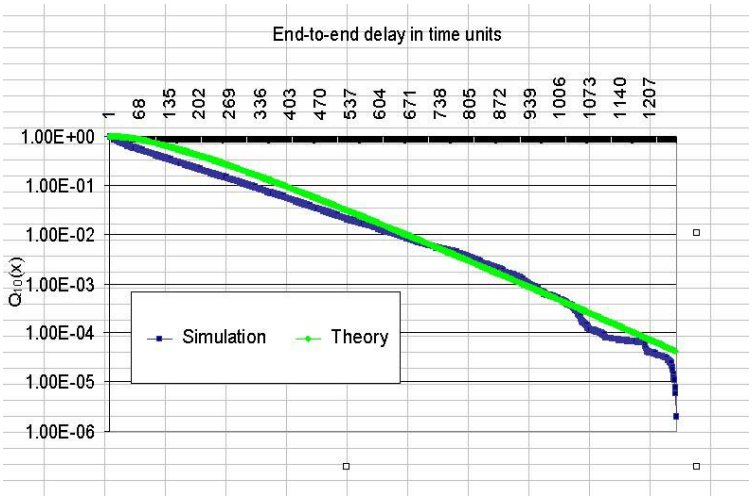

Figure 2 End-to-end delay distribution for a network of 10 buffers

This paper extends the analysis given in [1] by observing not only the voice traffic, but also the TCP-based data transfer. We believe this approach is necessary in the current heterogeneous Internet, where different traffic classes compete for the network resources. Dimensioning the network with respect to only the VoIP traffic could result in great losses in terms of QoS for the data traffic, in the case of high data traffic load.

3.1. Queueing delay for Data Traffic and VoIP traffic

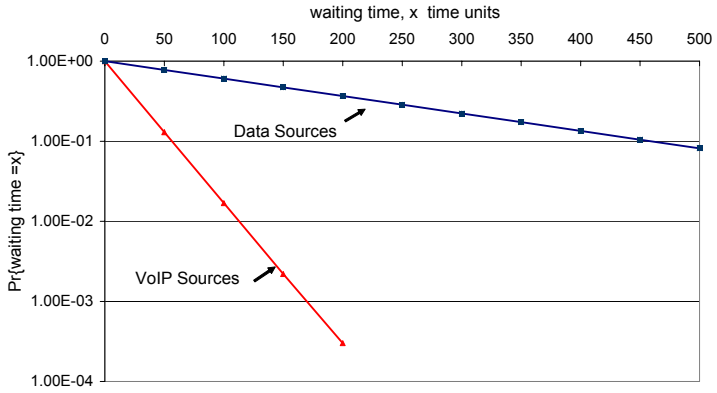

Figure 3 Comparison of the Waiting time distribution at a single buffer (FIFO Scheduling) for a multiplex of 10 traditional voice sources and 10 Data sources (cell-scale queueing omitted for clarity)

Figure 3 compares the waiting time distribution for a multiplex of traditional voice (on-off) source models and Data sources at a single buffer using FIFO scheduling. This is a result of a simulation in which both data and VoIP packets were 48 bytes long, and data traffic was not simulated as TCP traffic. These bursty sources are potentially the most difficult for networks to cope with. Result in Fig.3 shows that data sources suffer greater delays in a multiplexed buffer because of their burstiness. In this paper we extend this experiment by observing a more realistic Internet data model, with packet sizes of 1000 bytes for data, while the on-off values remain constant. We continue using the data from Table 1 to simulate VoIP traffic.

3.2. Dimensioning the multiservice IP Network – FIFO Queues

The problem of analytically tracking the queueing delay, processing delay and buffer size in a multi-hop, multi-service IP network has been widely documented. In order to examine in more detail the performance of multiple VoIP traffic flows in a realistic TCP/IP environment numerous simulation experiments were performed for this paper. The main objective of this analysis is to form a basic understanding of the impact of short-lived TCP traffic flows on VoIP performance. This understanding is very important for the future dimensioning of the IP networks that support voice services.

Figures 4 and 5 show the end-to-end delay for VoIP traffic and the average throughput of the background TCP data traffic. The observed cases include 10, 40 and 80 simultaneous VoIP traffic flows being active in the network. These three cases roughly correspond to the small, average and high number of active VoIP calls in the network. Overall link utilization (VoIP + TCP traffic) in the region of 0.4-0.8 has been observed. Two cases are examined:

(1) Data traffic and VoIP traffic are serviced concurrently, with a FIFO scheduler; (2) There is a differentiation of the traffic using CBQ/WRR scheduling.

Naturally, the average delay for VoIP traffic increases with the increased traffic load. While this is expected, it is important to dimension the traffic load at which delay average is below 200ms. Due to the statistical multiplexing gain, the increasing number of VoIP flows produces smaller average delays for the same link utilizations. In terms of the throughput, it is interesting to observe that higher number of active VoIP calls for the same link utilization results in higher TCP throughput. This is another consequence of the statistical multiplexing gain. Naturally, the evaluation of the TCP throughput is essential in the process of network dimensioning.

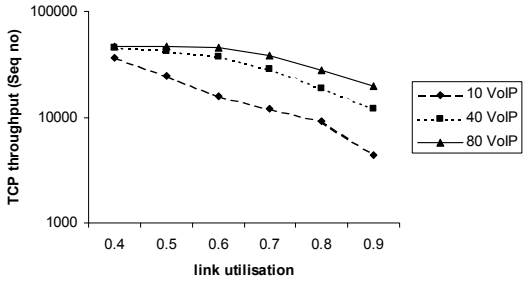

Figure 4 TCP throughput

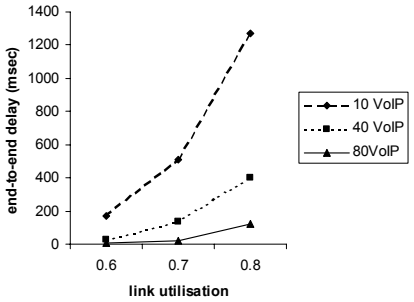

Figure 5 Comparison of end-to-end delay

Table 3 shows results of an experiment which shows what happens when an FTP application transfers a large file (10MB) end-to-end. In this experiment the link capacity was kept constant at utilisations 0.5 for 10 active VoIP calls, 0.7 for 40 VoIP calls and 0.8 for 80 VoIP calls. These utilizations follow traditional network dimensioning practice – systems with few users need to be dimensioned for peak values; systems with many users need to be dimensioned for average values [10].

All cases suffered badly during the file transfer. These results show that in the best-effort Internet, it is necessary to further decrease the utilization to accommodate the burstiness inherent in data traffic. The other option is to use deterministic techniques, such as traffic partitioning, in order to provide end-to-end delay guarantees for the VoIP traffic.

	Average VoIP packet delay	**Average VoIP packet delay during a large file transfer**
10 VoIP flows at 0.5 utilisation	38ms	390ms
40 VoIP flows at 0.7 utilisation	135ms	820ms
80 VoIP flows at 0.8 utilisation	125ms	650ms

Table 3.

3.3. Dimensioning the multiservice IP Network – Priority Queues

The most obvious solution for providing QoS guarantees to real-time traffic is strict partitioning of the network bandwidth. This would protect the real-time traffic from the burstiness of the data traffic. Such prioritized treatment of the traffic in the network is in the basis of a large proportion of QoS solutions proposed recently, including the IETF Integrated and Differentiated Services architectures.

The priority queuing works in the following way – several traffic classes are served at the same time, but each traffic class has a weight associated. This weight roughly defines the percentage of time the server will be serving the packets belonging to the class. The important issue in network dimensioning in the environment of priority queues is the dimensioning of these weights. In the simulations performed for this paper, we experimented with the traffic loads already analysed in section 3.2. The objective was to determine whether priority queuing with appropriate weight dimensioning can provide better overall results.

The main issue here is the trade-off between the benefits VoIP traffic gets through capacity partition and the drawbacks felt by the TCP traffic.

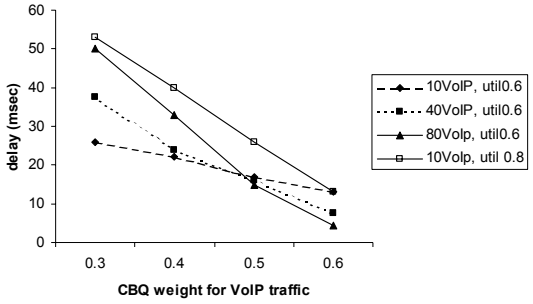

Figure 6 CBQ delay analysis

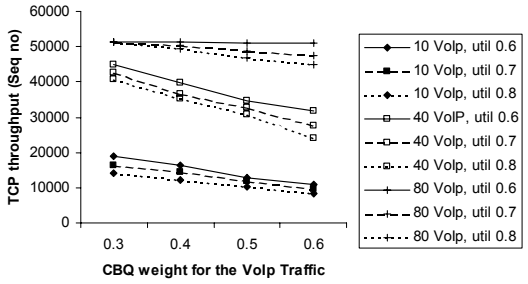

Figure 7 CBQ Throughput analysis

Fig 6 shows the results of experiments with CBQ schedulers. As the weight for VoIP traffic is decreasing from 0.6 to 0.3, we can see that the delay increases most in the case of large number of VoIP flows. However, CBQ scheduling provides excellent protection for VoIP traffic and the delays are all below the threshold of 200ms. This is not the case for 40 and 80 VoIP flows when traffic load is high (0.8). Our results show that the average delay for 40VoIP increases to 525ms when the weight of 0.3 is applied.

However, the introduction of capacity partitioning using priority schedulers decreases the performance of TCP traffic. As before, the larger number of VoIP calls (80) produces better results for the TCP traffic. It is noticeable from Fig.7 that increasing the space (by changing the CBQ weight) for the TCP traffic does not change the end-to-end performance a lot, at least in terms of the throughput. This is due to the ability of the TCP traffic to adapt to the available capacity and grab as much bandwidth as possible.

4. Conclusion

In this study we investigated further, resource-provisioning techniques based on end-to-end delay and loss characteristics. Previous work resulted in end-to-end provisioning guidelines in enterprise VoIP networks. This work uses simulation to monitor the impact (on end-to-end delay and throughput) of voice-data integration. The first network scenario investigated in this paper was based on a shared network environment with Voice and Data traffic scheduled using a FIFO scheduler. Results produced showed the impact of the statistical multiplexing gain inherent in multiplexing a larger number of VoIP calls and also helped to validate the simulation model.

The second part of the study examined the end-to-end delay and throughput behaviour of voice and data traffic using CBQ/WRR scheduling, which provides resource partitioning.

Results presented demonstrate the control provided by weight assignment as expected, and this is reflected in the end-to-end delay figures of the voice traffic. However, the impact of reducing the resources to the data traffic has less impact than might be expected.

There are several interesting areas to be further investigated. Assumptions have been made with regard to the parameters chosen for the model; nevertheless results obtained and analyzed provide a good foundation for the introduction of more complex traffic sources reflecting the actual hardware in use on VoIP networks. Our study was designed to provide an insight into the challenges of dimensioning and modeling large-scale voice-data integrated networks and to develop guidelines for service providers.

5. References

[1] R. Stewart, J. Schormans, "*Accurate Dimensioning of Service Level Agreements in Voice-over-IP Networks*", Journal of the Institution of British Telecommunications Engineers, Vol 2, part 3, July-September 2001

[2] S. Floyd, V. Jacobson, "*Link-Sharing and Resource Management Models for Packet Networks*", IEEE/ACM Trans. on Networking, Vol.3, No.4, August 1995

[3] D. Berstekas, R. Gallagher, "*Data Networks*", Prentice-Hall, 1992

[4] ns-2 Network Simulator, available at *http://www.isi.edu/nsnam/ns*

[5] R. Stewart, "*End-to-end Delay analysis for Small/Medium Scale IP netw*orks" PhD Thesis, University of London 2003

[6] D. Yates, J. Kurose, D. Towsley, M. G. Hluchyj, "*On per-session end-to-end delay distributions and the call admission problem for real-time applications with QoS requirements*", ACM SIGCOMM 93

[7] J.M.Pitts, J.Schormans, "*End-to-end QoS bounds for RTP-based service subnetworks*", SPIE Conference, Boston, USA, September 1999

[8] V.Rakocevic, J.Griffiths, G.Cope, "*Performance Analysis of Bandwidth Allocation Schemes in Multiservice IP Networks using Utility Functions",* International Teletraffic Congress, ITC17, Salvador da Bahia, Brazil, December 2001

[9] H. Awadalla, "Resource Management for Multimedia Traffic over ATM Broadband Satellite Networks, PhD Thesis, University of London, 2000

[10] O. Hersent, "IP Telephony: Packet-Based Multimedia Systems"