Optimal quality-level adaptation and performance evaluation for SLA-based VoIP services over DiffServ/MPLS networks

Hui-Kai Su

Department of Electrical Engineering,
National Formosa University,
Yunlin 632, Taiwan
Fax: +886 5 6315609
E-mail: hksu@nfu.edu.tw

Abstract: This work proposes an optimal quality level adaptation mechanism based on contracted service level agreement (SLAs) and defines an optimal service policy for a Voice over IP (VoIP) application provider over DiffServ/MPLS networks. A VoIP application provider between VoIP users and network service providers faces the challenge of SLA mapping to maximise its profit and satisfy users' requirements. The numerical results show that the proposed mechanism adapts the quality level of each active call to new network conditions well, and maximises the profit under the constraints on the contracted SLAs. The contributions would help application providers develop call admission control (CAC) schemes and proprietary service policies efficiently.

Keywords: voice over IP; VoIP; quality of service; QoS; service level agreement; SLA; DiffServ; MPLS; quality adaptation.

Reference to this paper should be made as follows: Su, H-K. (2012) 'Optimal quality-level adaptation and performance evaluation for SLA-based VoIP services over DiffServ/MPLS networks', *Int. J. Autonomous and Adaptive Communications Systems*, Vol. 5, No. 2, pp.166–177.

Biographical notes: Hui-Kai Su received the BS from I-Shou University, Taiwan, in 1999. He received the MS and the PhD from National Chung-Cheng University, in 2001 and 2006 respectively. He was an Assistant Professor in the Department of Computer Science and Information Engineering, Nanhua University, Taiwan, during 2006 and 2009. He joined the Department of Electrical Engineering, Formosa University, in the spring of 2009. Currently, he is an Assistant Professor in the Department of Electrical Engineering. His research interests include multimedia network applications, P2P network applications, IP/MPLS network survivability, network QoS control and management, embedded systems, etc.

1 Introduction

A service level agreement (SLA) is a formal contract of the relationship between a service provider and its customers. The SLA also specifies what the customer can expect from his service provider, including service quantity and service quality, i.e., the services provided by the service provider, and the penalties paid by the service provider if it cannot meet the committed goals. Many publications on SLAs can be found in Verma (2004); Bouillet et al. (2002) and Long-Tae et al. (2001) and references therein.

In a VoIP service environment, people can be divided into three roles, namely *VoIP users*, *VoIP application providers* and *network service providers*. The VoIP users are concerned about what VoIP service price they must pay and what user-level quality of service (QoS) they may receive. The VoIP application providers care about the balance of revenue and investment, and providing a good VoIP service to users. Similarly, the network service providers are concerned about the balance of revenue and investment to provide a good transport service. Thus, the VoIP users and the VoIP applications provider would sign session-level SLAs (S-SLAs), while the VoIP application providers and the network service providers would also sign network-level SLAs (N-SLAs) between the application and transport service domains.

VoIP application providers can partition the framework of SLA management into resource planning and dynamic on-demand processing. The resource evaluation and planning are not discussed in this paper, and interested readers are referred to Su et al. (2006) and Chen et al. (2007). For the dynamic processing, VoIP users first stipulate the acceptable quality profiles in S-SLAs with their application provider, and the VoIP application provider purchases the network transport service signed in N-SLAs to support the VoIP application. After contracting S-SLAs and N-SLAs, the VoIP application provider has to adapt the VoIP quality level to maximise the profit and fulfil user requirements in any network condition according to the purchased network resource.

This work proposes a SLA-based quality level adaptation mechanism, and defines an optimal service model for a VoIP application provider under a broad range of traffic conditions. We assume that the transport networks have deployed QoS technologies to provide QoS guarantee, i.e., DiffServ/MPLS networks. The problem of adapting the VoIP quality level in the dynamic on-demand processing is formulated as a binary linear programming problem. Finally, the numerical results show that the proposed mechanism adapts the quality level of each active call effectively to any new network condition, and maximises the profit for the VoIP application provider under the constraints of the contracted S-SLAs and N-SLAs.

The remaining part of this paper is organised as follows. In Section 2, we introduce the related works about SLA and QoS management. In Section 3, our system environment are presented, and the quality adaptation problem is addressed. Our optimal SLA-based quality adaptation mechanism is proposed in Section 4, and the performance evaluation is shown in Section 5. Finally, Section 6 concludes this paper.

2 Related works

A lot of works on the issues of SLA definition, architecture and SLA management can be found in Marilly et al. (2002); Lewis and Ray (1999); Furuya et al. (2004). The structure of SLA real-time management for multi-service packet networks and network admission control are discussed in Bouillet et al. (2002). From the view of business point, the maximising benefit under users' requirements for service overlay network (SON) operator is discussed in Tran and Ziegler (2007).

Another study that uses an utility model to formulate the adaptive QoS management problem and maximise the profit of a network service provider can be found in Long-Tae et al. (2001). To maximise the VoIP provider's utility, it dynamically adapts the operating quality of each VoIP session among a set of acceptable operating qualities under the resource constraint. Although the utility model contains illuminating discussions about QoS adaptation, it is driven by some prior off-line evaluation results. Consequently, it overlooks some fundamental questions. For example, it did not consider the influence of network condition, i.e., delay, loss and jitter, on VoIP operating quality, or the derivation of network-level resource requirements from user-level QoS requirements.

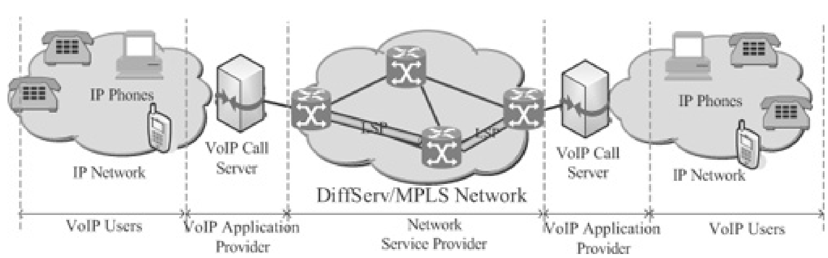
Based on a realistic VoIP service environment, this paper is to play the role of a VoIP service provider i.e. a SON operator. The SLA-based quality level adaptation problem for online phase (or operation phase) is extended from Su et al. (2006). Additionally, this problem is modelled to an utility model and formulated to an optimisation problem. The dynamic changes of network condition and call arrival/leave are considered in this paper. The contribution can help a VoIP service provider to maximise the total profit under the contracted S-SLAs and N-SLAs in any realistic environment condition (i.e. any VoIP traffic condition and any network condition).

3 System environment

Actually, VoIP service environment may be a mesh topology constructed with several VoIP call servers. To discuss easily, Figure 1 illustrates a part of the enterprise VoIP service environment. Unlike the conventional best-effort VoIP service on internet, the VoIP application provider provides a QoS-oriented VoIP service via the QoS-enabled DiffServ/MPLS network. The VoIP application provider provides several quality profiles for its customers, e.g. high and low quality video/audio calls. Additionally, the network service provider offers a DiffServ/MPLS network to provide differentiated service (DiffServ) or simple class of service (COS) and bandwidth reservation to the VoIP application provider. For instance, high-quality call traffic is aggregated into a virtual trunk, called an label switched path (LSP) in MPLS networks, and treated as the highest priority traffic in the whole network to satisfy the transport requirements. To reduce the call setup time, the DiffServ-enabled LSPs should be established by the network provider in advance after contracting N-SLAs. Finally, the VoIP traffic of each call is aggregated into the proper LSP using session classification (Su et al., 2004) or other technologies.

The VoIP application provider in this environment not only provides the connectivity and interworking of VoIP services as well as the general VoIP application providers, but also offers the QoS-oriented VoIP service by integrating voice over IP over MPLS (VoIPoMPLS) or voice over MPLS (VoMPLS) technologies. The infrastructure of VoIP call servers, such as SIP proxy, SIP registrar, H.323 gatekeeper, signaling gateway and media gateway, must be established. Additionally, the VoIP application provider knows the extent of the dedicated network resources. Therefore, he can periodically probe the network quality and help user agents to determine an appropriate quality level to any new network condition during the call proceeding and their conversation.

Figure 1 VoIP services over DiffServ/MPLS networks



Thus, to maximise the total profit and satisfy the contracted S-SLAs under the contracted N-SLAs, the VoIP application provider has to help the VoIP users to decide the proper quality level dynamically and adapt their quality level in a new VoIP traffic and network condition.

4 Optimal SLA-based quality adaptation mechanism

To provide optimal service policies, this study models the behaviour of the VoIP quality level adaptation in Figure 2. The internal mapping functions (F_1,F_2, F_3 and F_4) are defined as below.

- *Quality-to-resource mapping* ($F_1(i, k)$): the mechanism maps session operation quality to the resource requirement. It means how much kth class network resource is needed to provide ith quality call. The measurement results of transport network indicate that the VoIP application provider is aware of the network abilities, such as the quality statistics of each network DiffServ class in any particular network condition. Hence, the function F_1 can be defined by the VoIP application provider based on the quality statistics and the QoS requirements of each service quality. Briefly, F_1 is dynamically configured according to the measured network conditions.
- *Quality-to-revenue mapping* ($F_2(i)$): the mechanism maps session operating quality to the generated revenue. It indicates how much revenue is received by a ith quality call.
- *Violation-to-penalty mapping* ($F_3(i, j)$): the mechanism maps SLA violations to the incurred penalty. If the quality of jth call is downgraded to ith quality under the contracted S-SLA, the violation penalty is incurred.
- *Inefficiency-to-penalty mapping* ($F_4(k, L'_k/\eta_k R_k)$): the mechanism maps resource inefficiency to the incurred penalty. $L'_k/\eta_k R_k$ indicates the network target utilisation of kth-class network resource. Thus, F_4 function can control the network inefficiency penalty. Either the expensive network resource or the cheap network resource would be used first.

The VoIP application provider should configure the above internal mapping functions to enforce the service policy. The route selection policy plays a significant role in avoiding S-SLA violation, and is implied in F_1. Additionally, the VoIP application provider should also specify the target utilisation level of each network DiffServ class (η_k). To regulate the value of function F_1 in various network conditions, the network ability should be measured periodically. The accuracy of the network measurement is not discussed in this study; however, several network measurement techniques can be found in Hu and Steenkiste (2003); Ali and Lepage (2007); Tan et al. (2006) and Ekelin et al. (2006).

All parameters are listed in Table 1. Besides the above internal mapping functions and parameters, to decide the optimal service policy, the VoIP application provider has to give the following parameters: the set of acceptable operating qualities (quality profile: Q_i) of each call, the number of quality profiles (h), the amount of current active calls (n), the number of network DiffServ classes (m), the reserved bandwidth of each network service class (R_k), the target utilisation level of each network service class (η_k) and the average load effective ratio of each call (ρ) are acquired.

Figure 2 The utility model of VoIP quality level adaptation

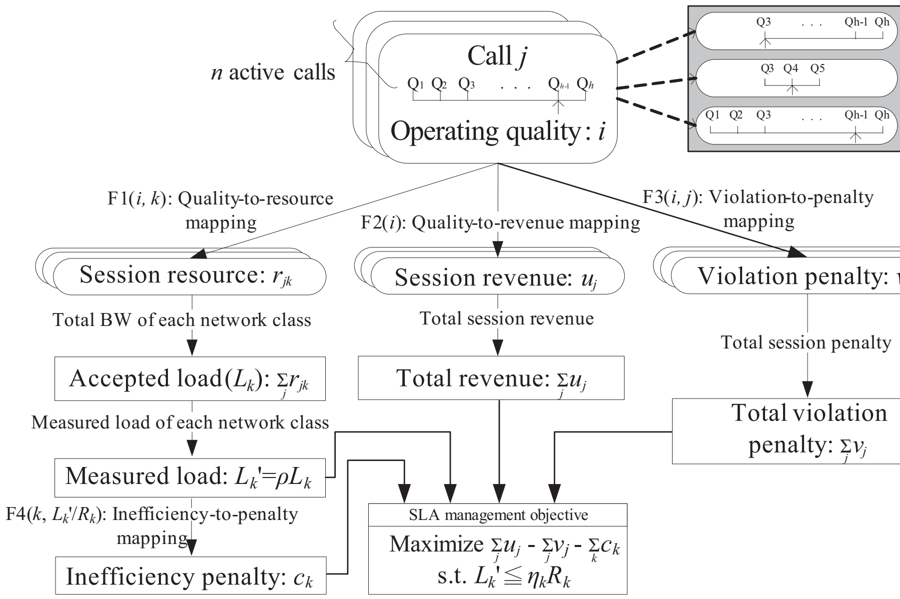


Table 1 Parameter list

Notation	*Description*
h	The number of quality profiles, e.g. 1st VoIP quality profile, 2nd VoIP quality profile, ..., hth VoIP quality profile
m	The number of network DiffServ, e.g. 1st network class, 2nd network class, ..., mth network class
n	The amount of current active calls, e.g. 1st call, 2nd call, ..., hth call
Q_i	The quality profile i where $i = 1, 2, \ldots, h$, e.g. $Q_1, Q_2, \ldots, Q_h$
R_k	The reserved bandwidth of each network service class where $k = 1, 2, \ldots, m$, e.g. $R_1, R_2, \ldots, R_m$
η_k	The target utilization level of network service class k where $k = 1, 2, \ldots, m$
ρ	The average load effective ratio of each call
u_j	The revenue of call j where $j = 1, 2, \ldots, n$
v_j	The violation penalty of call j where $j = 1, 2, \ldots, n$
c_k	The inefficiency penalty of network class k where $k = 1, 2, \ldots, m$
x_{ijk}	The decision variable. If $x_{ijk} = 1$, then call j is operated in the quality profile i and treated as the network class k. Otherwise, $x_{ijk} = 0$

To maximise the total profit, the appropriate quality of each call can be determined by solving the binary linear programming problem in Equations (1)–(4). Let u_j, v_j and c_k denote the revenue of call j, the violation penalty of call j and the inefficiency penalty of network service class k, respectively. The total profit is equal to the total revenue decreased by the total violation penalty and the network inefficiency penalty. Finally, the decision variable x_{ijk} can be resolved. If $x_{ijk} = 1$, the call j is operated in the quality profile i and treated as the network service class k. Otherwise, $x_{ijk} = 0$. The optimisation formulations are shown below.

Objective:

$$\text{Maximise } \sum_{j=1}^{n} u_j - \sum_{j=1}^{n} v_j - \sum_{k=1}^{m} c_k \tag{1}$$

s.t.

$$L'_k \leq \eta_k R_k, \quad k = 1, 2, \ldots, m \tag{2}$$

$$\sum_{k=1}^{m} \sum_{i=1}^{h} x_{ijk} = 1, \quad j = 1, 2, \ldots, n \tag{3}$$

$$x_{ijk} = 1/0, \quad i = 1, 2, \ldots, h, \quad j = 1, 2, \ldots, n \quad \text{and} \quad k = 1, 2, \ldots, m \tag{4}$$

where:

$$L'_k = \rho \sum_{j=1}^{n} r_{jk}$$

$$r_{jk} = \sum_{i=1}^{h} F_1(i, k) \times x_{ijk}$$

$$u_j = \sum_{k=1}^{m} \sum_{i=1}^{h} F_2(i) \times x_{ijk}$$

$$v_j = \sum_{k=1}^{m} \sum_{i=1}^{h} F_3(i, j) \times x_{ijk}$$

$$c_k = F_4(k, L'_k/\eta_k R_k)$$

To find the maximal profit, this problem can be resolved simply by brute-force search. However, the calculating complexity is $O((h \times m)^n)$. This is, a non-deterministic polynomial-time hard (NP-hard) problem. With the increasing amount of active VoIP calls, the computing time is very significant.

5 Performance evaluation

5.1 Numerical result

A VoIP application provider is assumed to provide a QoS-oriented VoIP service between both enterprise networks. In Table 3, the VoIP application provider provides three quality profiles ($h = 3$) to their customers, denoted as Q_1, Q_2 and Q_3. The customers contract these *three quality profiles* and accept *dynamic quality level adaptation* during communication. Additionally, the average call holding time is set to 3 min. The F_2 function is listed in Table 2. $F_2(Q_1)$, $F_2(Q_2)$ and $F_2(Q_3)$ of each 3-min call equal \$0.17, \$0.1 and \$0.06, respectively. After beforehand SLA evaluation, the VoIP application provider purchases the network bandwidths of three DissServ classes ($m = 3$), namely EF class, assured forwarding (AF1) class and AF2 class, where $R_1 = 768$ Kbps, $R_2 = 1280$ Kbps and $R_3 = 256$ Kbps.

Table 2 An example of F_2 function for 3-min call

Qualities (Q_i)	*Call revenue*
Gold (Q_1)	\$0.17
Silver (Q_2)	\$0.1
Copper (Q_3)	\$0.06

Table 3 An example of quality profiles and F_1 function

Qualities (Q_i)	*Network req.*	$F_I^L(i,k)$	$F_I^H(i,k)$
Gold (Q_1)	BW $\geq$ 384 Kbps	EF: 384 Kbps	EF: 384 Kbps
	delay $\leq$ 50 ms	AF1: 384 Kbps	AF1: ∞
	jitter $\leq$ 40 ms	AF2: ∞	AF2: ∞
	loss $\leq$ 1 %		
Silver (Q_2)	BW $\geq$ 256 Kbps	EF: 256 Kbps	EF: 256 Kbps
	delay $\leq$ 100 ms	AF1: 256 Kbps	AF1: 256 Kbps
	jitter $\leq$ 60 ms	AF2: 256 Kbps	AF2: ∞
	loss $\leq$ 3 %		
Copper (Q_3)	BW $\geq$ 128 Kbps	EF: 128 Kbps	EF: 128 Kbps
	delay $\leq$ 200 ms	AF1: 128 Kbps	AF1: 128 Kbps
	jitter $\leq$ 80 ms	AF2: 128 Kbps	AF2: 128 Kbps
	loss $\leq$ 5 %		

The network ability is assumed to be measured periodically by the VoIP application provider. Table 3 shows the F_1 functions of both network conditions, namely low network loading and high network loading. In the low network loading, the network requirements of the Q_1 VoIP service, in terms of bandwidth, delay, loss and jitter, can be satisfied by the EF class and AF1 class bandwidths (F_1^L). However, because the high network loading may have heavy network congestion, it can only be satisfied by the EF class bandwidth (F_1^H) in the high network loading.

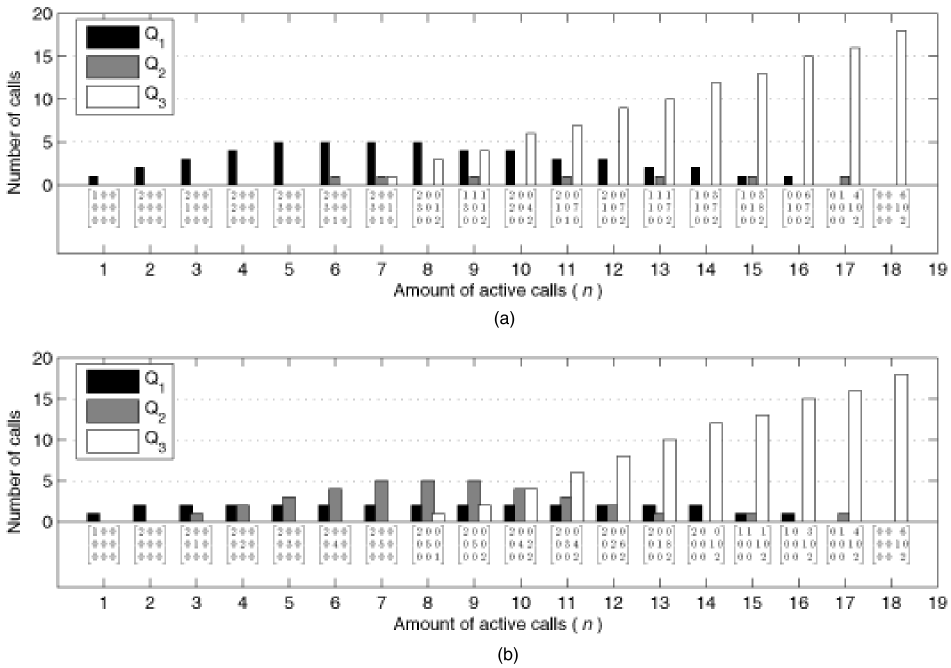
To increase the network usage of the expensive network resource, the network inefficiency penalty is considered. The F_4 function is listed in Table 4. $F_4(k, L'_k/\eta_k R_k)$ functions of EF class, AF1 class and AF2 class are defined as $(1 - L'_k/\eta_k R_k) \times 10^{-3}$, $(1 - L'_k/\eta_k R_k) \times 10^{-4}$ and $(1 - L'_k/\eta_k R_k) \times 10^{-5}$, respectively. ρ and η_k for each DiffServ class k are set to 1. Since the customers set all quality profiles and accept dynamic quality level adaptation, no SLA violation occurs here, and the F_3 function is set to 0.

The problem is calculated by a high-performance optimisation software (i.e. IBM ILOG CPLEX Optimizer IBM ILOG CPLEX Optimizer (2010)). Figure 3 (a) and (b) shows the quality level distributions of the active calls in low and high network loading conditions, individually. The matrix Y shown below the bars is defined as the optimal operation status. The y_{ik} calls are operated in Q_i, and transmitted at network class k. For instance, while $n = 7$, the optimal solution in the low network loading is $\#Q_1 = 5$, $\#Q_2 = 1$ and $\#Q_3 = 1$. Furthermore, two of the five Q_1 calls are transmitted at EF class, and the other three are transmitted at AF1 class.

Table 4 An example of F_4 function

Network class	*Inefficiency penalty*
EF ($k = 1$)	$(1 - L'_1/\eta_1 R_1) \times 10^{-3}$
AF1 ($k = 2$)	$(1 - L'_2/\eta_2 R_2) \times 10^{-4}$
AF2 ($k = 3$)	$(1 - L'_3/\eta_3 R_3) \times 10^{-5}$

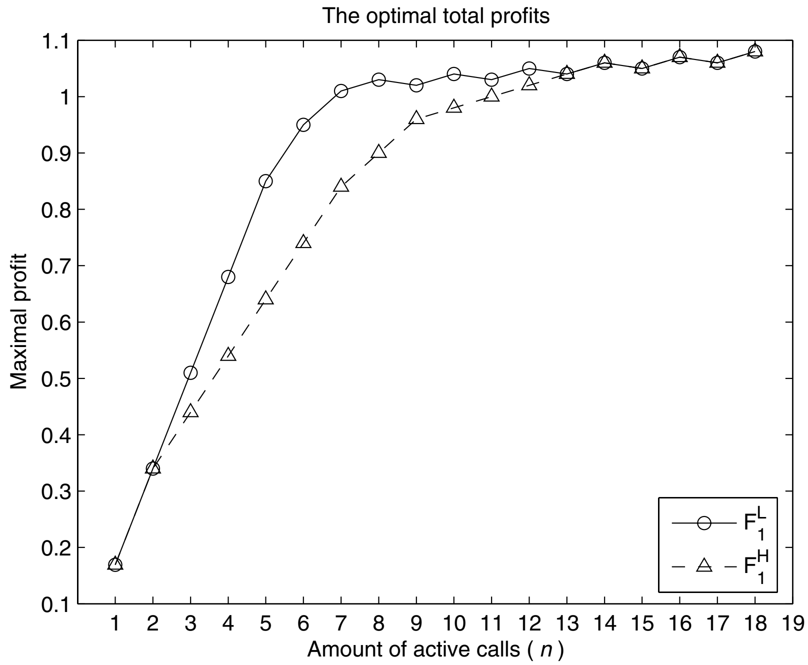
Figure 3 The quality level distributions



Experimental results demonstrate that each call with the optimal service policy can be successfully adapted to a new traffic and network condition. In the condition of low active calls and low network loading, the higher quality profile is selected first, and the calls are transmitted using the cheapest network bandwidth to maximise the total profit. For example, several Q_1 calls are treated as AF1 class between $n = 3$ and $n = 14$. However, in the condition of high network loading, because the F_1^H constraints are stricter than that of F_1^L, the number of Q_1 calls is limited, while the available EF class resource is empty. Additionally, regardless of network loading, the quality levels in low and high network loading conditions are degraded to maximise the total capacity when the number of active calls is increasing.

Figure 4 shows the total optimal profits (capacity = 18 calls). Since the constraints of the higher quality profile requirements are relaxed in the low network loading, the total profit in low network loading is observably higher than the profit in high network loading. Thus, the results would help the VoIP application provider develop call admission control (CAC) policies. The CAC should either degrade the quality levels of the existent active calls and accept the new call to maximise the capacity, or reject it to keep the maximal profit.

Figure 4 The optimal total profits (capacity = 18 calls)



5.2 Computing cost

To evaluate the computing cost, we follow the above scenario and extend the system scale to 10,000 active VoIP calls. We assumes that the VoIP application provider purchases the network bandwidths of three DiffServ classes ($m = 3$), namely EF class, AF1 class and AF2 class, where $R_1 = 256$ Mbps, $R_2 = 768$ Mbps and $R_3 = 256$ Mbps, respectively. When each call operates with the lowest quality level (Q_3), the maximal system capacity equals 10,000 calls.

Figure 5 shows the total optimal profits (capacity = 10,000 calls). The profit increasing trend is similar to Figure 4. Due to the loose constraints of the higher quality profile requirements in the low network loading, the total profit in low network loading is observably higher than the profit in high network loading. Additionally, the maximal gap of both curves falls in the point ($n = 4,000$). It means that the profit is very sensitive to network loading when the system loading equals to $4,000$ calls/$10,000$ calls, and the quality-level adaptation is very significant around this point.

The evaluation of our computing cost is illustrated in Figure 6. The results are calculated by a general personal computer, which is using Intel(R) Core(TM)2 Duo CPU E7400 @ 2.8 GHz and 3.46 G RAM. The mean of solving time is averaged out from 1,000 solving times. The trades of both curves are close to linearly increase. The point of $n = 7,000$ with F_1^H is a special state, the Optimiser software needs more time to solve this optimal problem. It may depend on its solving algorithm. Briefly, the solving time is acceptable, and the quality-level adaptation of each call can be performed periodically.

Figure 5 The optimal total profits (capacity = 10,000 calls)

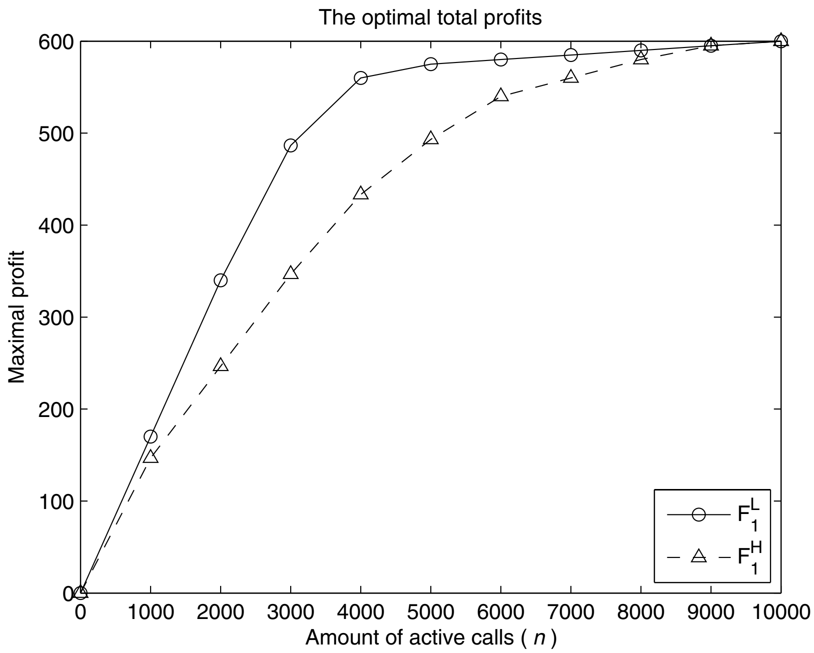
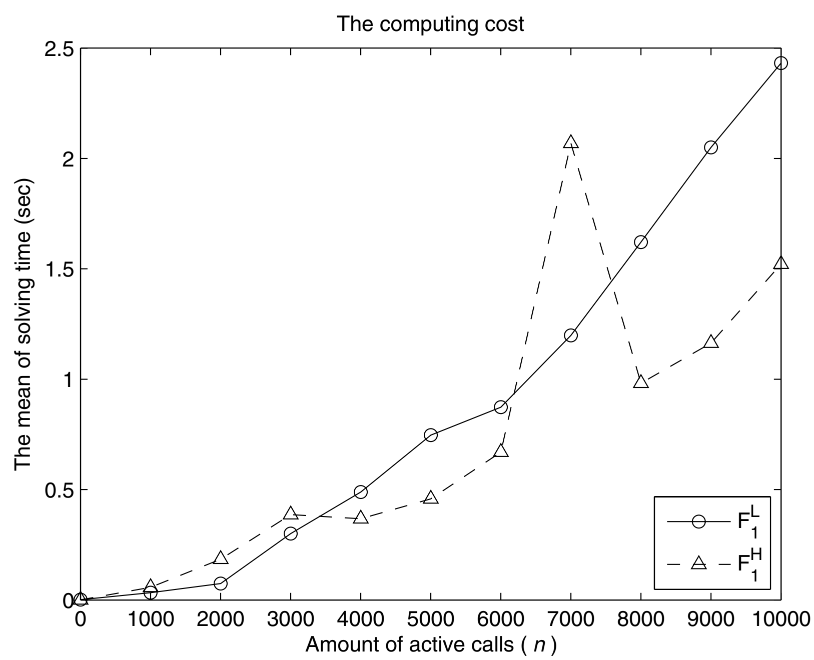


Figure 6 The computing cost (capacity = 10,000 calls)



6 Conclusion

This work proposes an optimal SLA-based quality adaptation mechanism, and defines an optimal service model for a VoIP application provider over DiffServ/MPLS networks between two VoIP call servers. The problem of the VoIP quality level adaptation is formulated in the dynamic on-demand processing as a binary linear programming problem. The numerical results show that the proposed mechanism can adapt the quality level of each active call to the variety of traffic and network conditions well, and maximise the total profit under the constraints of the contracted S-SLAs and N-SLAs. Additionally, the solving time of the optimisation model is acceptable to our system. Thus, such contributions would help a VoIP applications provider to develop CAC and proprietary service policies efficiently. In the future, we are also going to develop a heuristic mechanism to improve the calculating speed for a large-scale environment.

Acknowledgements

We would like to thank the National Science Council (NSC) in Taiwan. This research was supported in part by the NSC under the Grant Number NSC-98-2218-E-150-006 and NSC97-2221-E-150-079.

References

Ali, A. and Lepage, F. (2007) 'IGMPS, a new tool for estimating end-to-end available bandwidth in ip network paths', *Proceedings of the IEEE ICNS*.

Bouillet, E., Mitra, D. and Ramakrishnan, K. (2002) 'The structure and management of service level agreements in networks', *IEEE Journal on Selected Areas in Communications*, Vol. 20, pp.691–699.

Chen, H., Su, H-K. and Cheng, B-C. (2007) 'An objective-oriented service model for VoIP overlay networks over DiffServ/MPLS networks', *Computer Communications*, Vol. 30, pp.3055–3062.

Ekelin, S., Nilsson, M., Hartikainen, E., Johnsson, A., Mangs, J-E., Melander, B. and Bjorkman, M. (2006) 'Real-time measurement of end-to-end available bandwidth using kalman filtering', *Proceedings of IEEE/IFIP NOMS*.

Furuya, H., Nomoto, S., Yamada, H., Fukumoto, N. and Sugaya, F. (2004) 'Toward QoS management of VoIP: experimental investigation of the relations between IP network performances and VoIP speech quality', *IEICE Transactions Communication*, Vol. E87-B, pp.1610–1622.

Hu, N. and Steenkiste, P. (2003) 'Evaluation and characterization of available bandwidth probing techniques', *IEEE Journal on Selected Areas in Communications*, Vol. 21, pp.879–894.

IBM ILOG CPLEX Optimizer (2010) Available at: http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/.

Long-Tae, P., Jong-Wook, B. and Hong, J.W-K. (2001) 'Management of service level agreements for multimedia internet service using a utility model', *IEEE Communications Magazine*, Vol. 39, pp.100–106.

Lewis, L. and Ray, P. (1999) 'Service level management definition, architecture, and research challenges', *Proceedings of IEEE GLOBECOM*.

Marilly, E., Martinot, O., Betge-Brezetz, S. and Delegue, G. (2002) 'Requirements for service level agreement management', *Proceedings of IEEE Workshop on IP Operations and Management*.

Su, H-K., Wu, C-S. and Chen, K-J. (2004) 'Session classification for traffic aggregation', *Proceedings of IEEE ICC*.

Su, H-K., Yau, Z-Z., Wu, C-S. and Chen, K-J. (2006) 'Session-level and network-level SLA structures and VoIP service policy over diffserv-based MPLS networks', *IEICE Transactions Communications*, Vol. E89-B, pp.383–392.

Tan, W., Zhanikeev, M. and Tanaka, Y. (2006) 'Abshoot: a reliable and efficient scheme for end-to-end available bandwidth measurement', *Proceedings IEEE TENCON*.

Tran, H.T. and Ziegler, T. (2007) 'A design framework towards the profitable operation of service overlay networks', *Computer Networks*, Vol. 51, pp.94–113.

Verma, D. (2004) 'Service level agreements on IP networks', *Proceedings of the IEEE*, Vol. 92, pp.1382–1388.