PAPER Special Section on Multimedia QoS Evaluation and Management Technologies

Session-Level and Network-Level SLA Structures and VoIP Service Policy over DiffServ-Based MPLS Networks

Hui-Kai SU[†a)], ***Student Member***, **Zhi-Zhen YAU**[†], **Cheng-Shong WU**[†], ***and*** **Kim-Joan CHEN**[†], ***Nonmembers***

SUMMARY This paper proposes a framework for session-level SLA (Service Level Agreement) and network-level SLA management to provide QoS-oriented application services over DiffServ/MPLS networks. DiffServ and MPLS technologies enhance the capability of QoS guarantee on IP network, and application service provider can provide QoS-oriented application services to their customers based on the transport networks. The example of using our approaches in the VoIP service involving the network service provider, the VoIP service provider, and the VoIP customers are examined. The session-level SLA between VoIP service provider and VoIP customer and the network-level SLA (N-SLA) between network provider and VoIP service provider are constructed in this paper. Besides, the VoIP service provider faces the QoS-mapping issue and the balance between revenue and cost, i.e., how to contract the N-SLA. Therefore, we introduce an off-line SLA evaluation scheme, a heuristic optimization algorithm and an on-line SLA process method to provide VoIP service policy, and then the optimal QoS-mapping can be resolved. The concept of this framework of session-level SLA and network-level SLA management can be extended easily into other real-time multimedia and non-real time data services.

key words: *SLA, QoS management, service policy, VoIP, MPLS, DiffServ*

1. Introduction

A service level agreement (SLA) is a formal contract of the relationship that exists between a service provider and his customers. It is also used to specify what the customer could expect from the provider, including service quantity and service quality, i.e., what services the service provider will furnish and what penalties the service provider will pay if he cannot meet the committed goals. A SLA can be contracted either dynamically or statically. Dynamic SLA could be negotiated by the service level negotiation protocol [1], [2] to contract automatically in order to satisfy user's demands dynamically. Static SLA contract is made between two human parties and its terms cannot be changed without human intervention. Moreover, a SLA could be written in either formal language like SLAng (Service Level Agreement Language) [3] or other format as long as both parties involved understand the content of a SLA completely.

A SLA is not valuable in itself, if it cannot be managed efficiently. To guarantee that the delivered service quality levels conform to the SLA contract, an efficient service level management or SLA management scheme is necessary. Many publications on service level agreements (SLA) can be found in various articles [4]–[6]. One of the most important trends is that the purpose of a SLA has recently shifted from a financial contract towards a process for the management of the delivered service quality levels. In addition, with the development of high-speed network and QoS control technologies, the SLA management issue in Internet is concerned by service provider and his customer as well as QoS guarantee.

Today, people on Internet can be divided into three roles: users, application providers and network service providers. Users are concerned about application price which they have to pay and user-level QoS (Quality of Service) which they may receive, e.g., call blocking rate and voice quality in VoIP (Voice over IP) service. Application providers care about the balance of revenue and investment, and how to provide a wonderful application service to users. Similarly, network providers are concerned about the balance of revenue and investment, and how to provide a good transport service. How to satisfy the expectations of each role depends on a suitable service contract and a fair and flexible QoS management framework. In fact, although many operators who are offering application services to the end users are often offering Internet access service at the same time, the operators also need a general framework of SLA management to resolve the QoS-mapping issue and the balances of application service and network service between revenue and cost.

In this paper, we introduce general enterprise environments for VoIP service over a QoS-based Internet, and analyze the relationships between session-level SLA and network-level SLA. The three major roles are VoIP user, VoIP service provider, and network service provider shown in Fig. 1. The VoIP service provider provides voice, videophone or videoconference services, and they may interwork with other VoIP service providers. In Fig. 1, two VoIP users communicate with each other, and they may or may not belong to the same VoIP provider. The VoIP users and their VoIP service provider should contract a vertical SLA defined as a session-level SLA (S-SLA) in this paper, where the session is defined as a lasting relationship or connection between a user (or user agent) and a peer on application layer, i.e. a call. Furthermore, both VoIP users in Fig. 1 have a horizontal relationship such as a family or a friend. If the both users communicate via different VoIP service providers, the horizontal SLAs also should be existed between the VoIP service provides. Similarly, VoIP service provider and the network service provider may also contract a vertical SLA defined as a network-level SLA (N-SLA) in this paper be-

Manuscript received May 2, 2005.
Manuscript revised August 12, 2005.

[†]The authors are with the Department of Electrical Engineering, National Chung-Cheng University, Taiwan, R.O.C.
a) E-mail: pat@ee.ccu.edu.tw

DOI: 10.1093/ietcom/e89–b.2.383

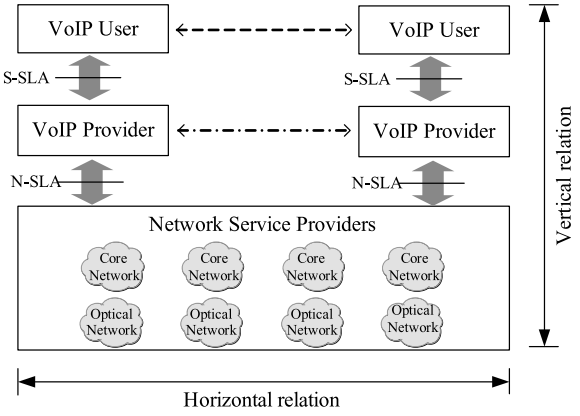


Fig. 1 SLAs between VoIP users, VoIP service provider and network provider.

tween VoIP service domain and transport service domain. The horizontal SLAs should be existed between the network service provides. Additionally, the transport network may include network providers in different OSI layer (e.g., core MPLS network and optical network providers), and the relationship or collaboration between them falls out of the interest of this work. Discussion on the more complex SLA environments can be found in [7], [8]. In the service framework, the VoIP service provider faces the QoS-mapping issue and the balance between revenue and cost, i.e., how to contract the S-SLA and the N-SLA and to bring maximum profit respectively.

In this paper, we propose a robust, fair, and efficient framework for session-level SLA and network-level SLA management to provide QoS-oriented application services over DiffServ/MPLS networks. We assume that QoS requirements are demanded by the VoIP users and the transport networks deploy DiffServ/MPLS technologies and have the capability of QoS guarantee. We play a role of VoIP service provider, and the QoS-mapping issue and the balance between revenue and cost are formulated to an optimal problem, which can maximize the profit of the VoIP service provider while satisfying S-SLA under revenue-to-cost ratio constraints. In order to find an optimal balance point efficiently between revenues and costs, we propose a heuristic optimization algorithm to decide the VoIP service policy, and then solve the optimal QoS-mapping problem.

The remaining part of this paper is organized as follows. In Sect. 2, we introduce the related works about SLA and SLA management. In Sect. 3, our system environments are presented. Our SLA management framework and SLA structure are proposed in Sect. 4 and Sect. 5 respectively. The SLA management and optimal VoIP service policy, including off-line QoS evaluation and on-line adaptive QoS tuning, is explained in Sect. 6. Finally, Sect. 7 concludes this paper.

2. Related Work

A lot of researches on the issues of SLA definition, architecture, and SLA management can be found in [9]–[11]. A framework for SLA management in a top-down approach can be found in [12], and the structure of SLA real-time management in multi-service packet networks and network admission control are discussed in [5]. In this paper, we consider the realistic environment. We try to play the role of VoIP service provider, propose templates of S-SLA and N-SLA for VoIP service over DiffServ/MPLS networks, and solve the QoS mapping issues and revenues between them.

Another works that use the utility model to formulate the adaptive QoS management problem and maximize the profit of the network service provider can be found in [6], [13]. The utility model is based on the concepts of QoS adaptation of VoIP sessions. To maximize 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.

In Sect. 5, we constructed a S-SLA to reflect the full user-level requirements for VoIP sessions. Thus, the resource requirements specified in N-SLA can be determined to fulfill the user requirements. The objective of the mechanism is to maximize the service profit for a VoIP service provider.

3. System Environment

The design of our SLA management framework and QoS mapping scheme is based on the system environments shown in Fig. 2. Unlike the conventional best-effort VoIP service on Internet, the VoIP service provider provides QoS-oriented VoIP service via QoS-enabled DiffServ/MPLS networks. In addition, the QoS-enabled DiffServ/MPLS networks may be composed of several AS domains. In order to reduce call setup time, we assume that the DiffServ-enabled LSPs are established by the network provider in advance after contracting N-SLAs. The VoIP traffic of each call would be aggregated into the proper LSP by session classification [14] or other technologies.

The VoIP service providers provide not only connectivity and interworking call services as well as the current VoIP service provider, but QoS-oriented VoIP service by integrating VoIPoMPLS (Voice over IP over MPLS) or VoMPLS (Voice over MPLS) technology. The infrastructure of VoIP call server and VoIP gateway, such as SIP proxy, SIP registrar, H.323 gatekeeper, signaling gateway or media gateway, should be established. Besides, because the dedicated network resource is known by the VoIP service provider, he should help the call endpoint to decide a proper service level to the new call during the call proceeding, e.g., an audio phone with low encoding rate will be selected while the

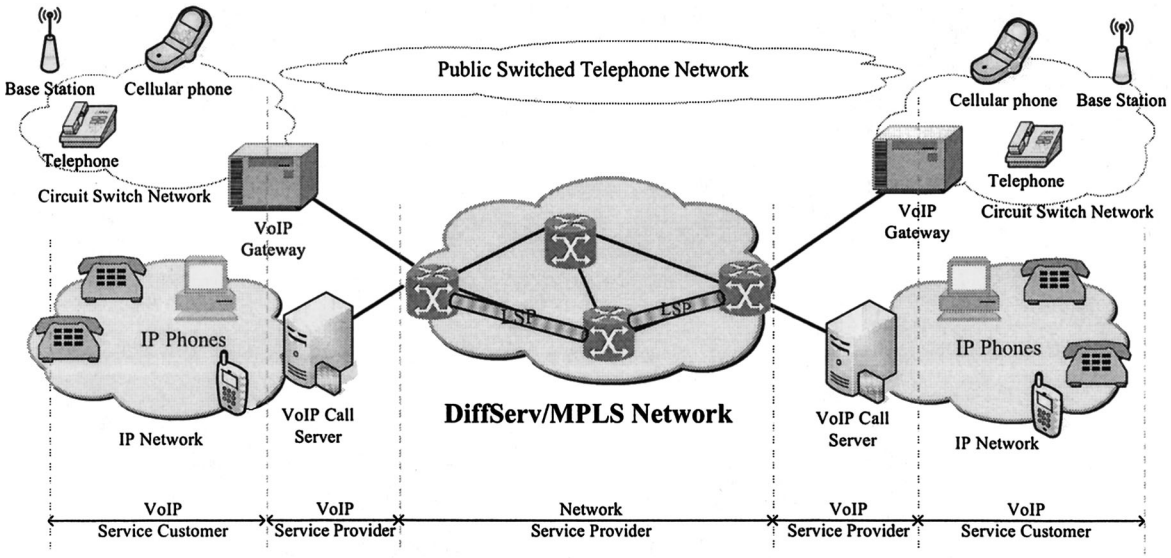


Fig. 2 QoS-oriented VoIP service over DiffServ/MPLS network.

network resource is poor or the network quality will be degraded.

The VoIP service customers are composed of two kinds of users. One is IP phone on IP networks, and the other is telephone on circuit-switch networks, e.g., traditional telephone, cellular telephone, ISDN telephone, etc. Although the telephone are the customer of circuit-switch network provider, they can buy the VoIP service and only pay the local communication cost to communicate to other foreign telephones with long distance over the DiffServ/MPLS networks via the VoIP gateways. However, the telephones include ISDN telephone can also communicate with each other directly over PSTN.

In the architecture, the S-SLA and N-SLA should be contracted, so the VoIP service provider faces the challenge of QoS-mapping issue and the balance between revenue and cost respectively, i.e., how to contract the N-SLA. In the below sections, they will be presented.

4. SLA Management Framework

In order to commit the contract, SLA is usually contracted loosely to tolerate the variation of system condition; however, some application required critical QoS cannot be compromised. Therefore, the issue of how to develop a SLA management framework to provide SLA predesign, optimal QoS mapping and on-line adaptive QoS tuning is very important. Three major concerns for such systems are performance, availability, and security. Performance requirements imply that these systems must be adaptable and self-configurable to the changes of workload. Availability and security requirements suggest that these systems also must adapt and reconfigure themselves to withstand attacks and failures.

We designed a robust, fair, and efficient SLA management model that continuously monitors the system workload and determines the optimal configuration to reach the goal of service provider in Fig. 3. In the SLA management model, the effective and fair QoS mapping which can reach the expectations of customers and service providers can be decided. The SLA metric can be kept per customer basis to achieve integrity. Additionally, the proposed SLA management is robust since it is adaptable to system condition changes, e.g., system loading and network loading.

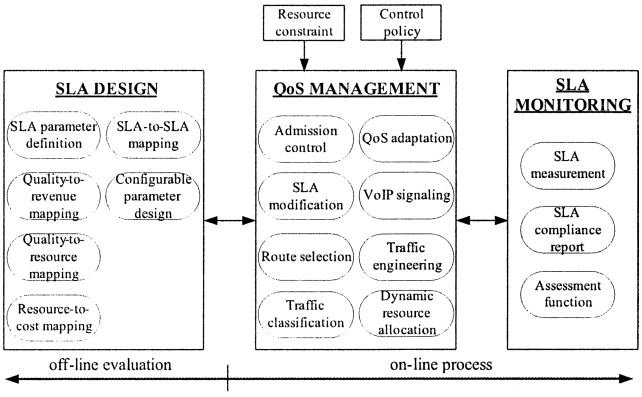


Fig. 3 SLA management framework.

In Fig. 3, the SLA management model contains off-line evaluation and on-line process and is explained as below. In the off-line evaluation phase, the issues of SLA pre-design and pre-plan are included, e.g., QoS mapping, SLA structure design, SLA negotiation and service deployment. The SLA is designed before services are provided; for instance, what SLA parameters are and how to maximize the profit of service provider. In addition, it also deals with the design of configurable parameter for service policy, such as SLA mapping constraints in the off-line phase and the high watermark for resource utilization in the on-line phase. More details will be clarified in Sect. 6.1.

On-line process intends to maximize the profit of service provider under the QoS mapping constraints while the system condition changes. It is composed with two different elements: QoS management and SLA monitoring.

For the assurance, the delivered service quality against SLA commitments should be monitored and measured to report customer about the real QoS that they received. SLA monitoring includes the SLA measurement, SLA compliance report, and the assessment function [15]–[17]. It interacts and cooperates with "QoS management" to make SLA management model adaptable to system condition changes.

- *SLA measurement:* The goal of SLA measurement is to measure the performance of a service, measure the delivered service qualities, and estimate the usage of various resources. It measures some significant metrics of QoS at specifies intervals and stores the collected results in its databases.
- *Assessment function:* The assessment function can ensure an early detection of QoS degradation, SLA violation, and illegal attack. It collects and inspects information on the monitored data to evaluate whether the metrics complies with configured watermarks and react to the events of QoS degradation, SLA violation, and illegal attack. The definition of events and how it reacts to the events could be configured through the manager of SLAs. For example, when SLA violation occurs, it could send notification report to the interested parties. Furthermore, it could adapt and reconfigure itself to withstand the events.
- *SLA compliance report:* SLA compliance report is sent to parties who are interested in the information and subscribe for it. The report is typically describing the compliance of delivered service qualities. People with different identities can subscribe different kinds of reports according to the authorities. Moreover, different kinds of reports are expressed as different kinds of format. For example, the SLA violation report could be sent to a SLA manager with critical format to take the necessary actions, such as policy change, QoS restoration, but sent to the customers with tender words for notification.

The goal of QoS management is to manage and control the delivered QoS to ensure compliance with the contracted SLA. We integrate the functionalities of network-level and session-level QoS management to compose the QoS management element, e.g., network capability and session admission control. QoS management is the core component in the proposed SLA management model, and it cooperates with SLA monitoring and SLA manager to provide the optimal VoIP service policy. In other words, the control parameters of QoS management may be configured manually by the manager of SLAs or tuned automatically by the assessment function of "SLA monitoring." In our project, both QoS managements are design and described below.

- *Network-level QoS management:* Network-level QoS management deals with underlying network capabilities. For instance, it includes connection admission control, route selection, traffic classification, resource reservation, dynamic resource allocation, buffer management, packet scheduling, etc. In practice, the most advanced models for QoS deployment, such as DiffServ and MPLS, still have to work together to provide full network-level QoS management, i.e., DiffServ-based MPLS network. Researches about DiffServ/MPLS network can be found in [18]–[20], the issues about implementation details fall out of the interest of this work. However, we can briefly mention that SLA service policies can be designed by combining the functionalities of DiffServ/MPLS QoS management with VoIP signaling management, e.g., SIP.
- *Session-level QoS management:* Session-level QoS management deals with the service policy to satisfy user-level QoS requirement specified in S-SLA. For instance, it deals with how to contract proper N-SLA, provide session admission control, adapt session operating qualities (QoS Adaptation), validate and modify SLA, etc. More details will be clarified in Sect. 6.2.

5. SLA Structures

Both S-SLA and N-SLA are involved in the SLA management of VoIP service provider and explained in the below subsections.

5.1 S-SLA Structure

We propose a wide range of different VoIP service classes to VoIP users, and different prices will be paid for each call class by the users. First, an enterprise or a group of VoIP users has to contract the S-SLA with the VoIP service provider and pay the regular fee to him. Second, according to the S-SLA, the VoIP service provider will propose an optimal service policy and contract the N-SLA with the network provider.

Besides the billing and discounting, the S-SLA structure is composed by quantity and quality parts. The quantity part of S-SLA depends on the behavior of VoIP users, and the quality part of S-SLA depends on the multimedia codec, e.g. coding method, coding rate, etc. In telecommunication networks, the characteristics of call arrival and call holding time approximate to Poisson distribution and Exponential distribution respectively. We assume the behavior of VoIP users is similar to that of the telephone users. According to Erlang-B formula, the VoIP provider can obtain the multiplexing gain of VoIP services as well as telecommunication providers. However, the SLA management framework may be applied to other real-time multimedia services other than VoIP service at the same time, e.g. video streaming service. Because the Poisson arrival is assumed for VoIP as well as the telephone, the evaluation of S-SLA quantity part for other non-VoIP services should be reconsidered according to the user's behavior and traffic pattern.

The quantity part contains *Busy Hour Traffic* and *Session Blocking Probability*. We assume that arrival of calls follows Poisson distribution with the rate of λ, and the call holding times are exponentially distributed with the mean holding time of $1/\mu$. The *Erlang* density is defined as a traffic load of VoIP calls, where $Erlang = \lambda/\mu$.

In the quality part, the VoIP provider provides h service levels (Q_i), where $1 \le i \le h$; for example, Q_1 expresses a low quality phone with low audio encoding rate, and Q_h expresses a high quality video phone with high video and audio encoding rates. Besides, the customers can indicate how many calls will be treated as Q_i at least. In the current VoIP signaling such as SIP and H.323, the operating quality is decided by the end users before the call connection established. However, the VoIP call servers in our system may intervene to select the proper operating quality during the call setup. For example, the VoIP call servers may snoop and modify the SDP (Session Description Protocol) contents in SIP messages to assign an proper audio codecs and a proper video codec for this call. Therefore, when a new call arrives, the operating quality of the call will be decided by VoIP users if the contracted quota of the selected operating quality is available, i.e. the amount of calls operating in the selected operating quality n_i is less than N_i. Otherwise, the proper operating service of the call will be dynamically and compulsorily decided during the call proceeding by the VoIP call server.

In the view of VoIP providers, they want to reduce the number of S-SLA parameters; otherwise users want to obtain a lot of guarantees in the view of VoIP users. However, if there is no enough bandwidth to transport user data, the transport quality will not be guaranteed. If there is enough bandwidth, it's the duty of VoIP providers to declare which quality profile they can provide. Therefore, all parameters of the quantity part and the quality profile are basic parameters and needed for S-SLA. Besides, the full S-SLA structure is explained as below.

- *Busy Hour Traffic* (*BHT*): The expected total traffic intensity of all quality profiles during the busiest hour of the day, e.g., 84 *Erlangs*.
- *Session Blocking Probability* (SBP): The probability of incoming calls failed during the Busy Hour, e.g., 0.1% SBP in one month average.
- *Callee parties*: A group of specified callee parties, e.g., VoIP service in an enterprise VPN. If customers specify that they want to communicate with public telephones, the VoIP service provider has to evaluate the traffic engineering for the customers and the result of the evaluation should be translated to the $FECset$ field in N-SLA.
- *Quality Profile* (Q_i): Quality profile is a set of acceptable operating qualities, i.e., Q_i, where $1 \le i \le h$. Q_1 is the lowest operating quality and Q_h is the highest operating quality.
- *Guaranteed Quantities* (N_i): The number of sessions operating in Q_i that VoIP provider guaranteed (while current traffic load not exceeding the BHT specified in SLA), i.e., N_i, where $1 \le i \le h$. N_1 is the guaranteed number of lowest quality sessions and N_h is the guaranteed number of highest quality sessions.
- *User Options*: Customers are allowed to submit their options for desired quality levels and bid prices during call proceeding, e.g., above the second lowest operating quality.
- *Billing Mechanism*: Billing mechanism for each operating quality, e.g., the communication rate for highest operating quality is 0.13 yens/second.
- *Regular Fee*: Regular payments from the VoIP users, e.g., Monthly fee is 100 yens.
- *Discounting policy*: The agree payments from the VoIP provider while the SLA commitments are invalidated, e.g., discount in case of service level degradation.
- *Others*: Other parameters of session-level contract, e.g., service level availability, etc.

5.2 N-SLA Structure

After contracting the S-SLA, the VoIP service provider has to decide how to contract the N-SLA with the network provider. The N-SLA structure also is separated into both categories of quantity and quality parts. The quantity part refers to the reserved bandwidth of each PHB (Per Hop Behavior) class, and the quality part, such as packet delay, lost and jitter, is implied in the PHB classes. Besides, the full N-SLA structure is described as below.

- *Total Bandwidth* (R): Total amount of resource required regardless of the network service classes, e.g., $R = 25.6$ Mbps.
- *Per Class Resource Guarantee* (R_k): Resource guaranteed for each network service class, i.e., R_k, $1 \le k \le n$. R_1 is the resource requirement of lowest priority class and R_n is the resource requirement of highest priority class.
- *FEC (Forwarding Equivalence Class) set*: a group of IP packets which are forwarded in the same manner, e.g., over the same path, with the same forwarding treatment. In other words, FEC set generally decides which packet will be delivered in the same MPLS LSP (Label Switched Path) by multi-field classification rules (source IP address, destination IP address, protocol ID, source port and destination port), e.g., the destination prefix of the callee parties are equal to "140.123.109.0/24."
- *Regular fee*: Regular payments from network customers, e.g., monthly fee is 800 yens.
- *Others*: Other parameters of network-level contract, e.g., network availability, etc.

6. SLA Management

6.1 Off-Line QoS Evaluation

Recall that from the VoIP provider's viewpoint, the S-SLA

and the N-SLA brings revenues and costs respectively. What should VoIP provider do after signing the S-SLA? With VoIP users' requirements in hand, the next step is undoubtedly to reserve proper resource to satisfy it. A rash VoIP provider may overestimate the resource requirement and waste a lot of money for it, while a miserly VoIP provider could underestimate it and cause a lot of complaints or penalties. Therefore, the VoIP service provider faces the translation issue between SLAs and the balance between revenue and cost. We introduce SLA management mechanism to reserve proper network resource in the phase of off-line evaluation and provide an optimal service policy in the phase of on-line process. Additionally, to accommodate unexpected user behaviors or network conditions, the worst case of system conditions are considered in the off-line evaluation, and then the operating quality of each call is adapted to the new conditions in the on-line process.

6.1.1 SLAs Translation

In Fig. 4, the reference model of SLA translation is illustrated. The goal of the problem is to maximize the profit of VoIP provider under the QoS and translation constraints, e.g., QoS requirements, network ability, revenue-to-cost ratio and maximum resource cost. Finally, the optimal N-SLA template will be produced.

In order to evaluate the network requirements, the VoIP service provider has to define internal QoS translation functions. Generally, the functions are the private assets of each service provider. However, the internal QoS translation functions are defined as below.

- *Quality-to-resource mapping* (F_1): The mechanism maps VoIP session operating quality to the resource requirement, e.g., the resource required by Q_1 and Q_2 are 32 kbps in AF class and 128 kbps in EF class, respectively. According to the network QoS measurements, the VoIP provider is aware of the network abilities, such as the quality statistics of each PHB class in a certain network condition. Therefore, the VoIP service provider can define F_1 function according to the quality statistics and the QoS requirements of each VoIP quality Q_i.
- *Quality-to-revenue mapping* (F_2): The mechanism maps VoIP session operating quality to the generated revenue, e.g., the revenues generated by Q_1 and Q_2 are 1 yen and 3 yens respectively.
- *Resource-to-cost mapping* (F_3): The mechanism maps operating resource requirement to the incurred network resource cost, e.g., the cost incurred by Q_1's resource requirement is 1.5 yens and incurred by Q_2's resource requirement is 13 yens, respectively.

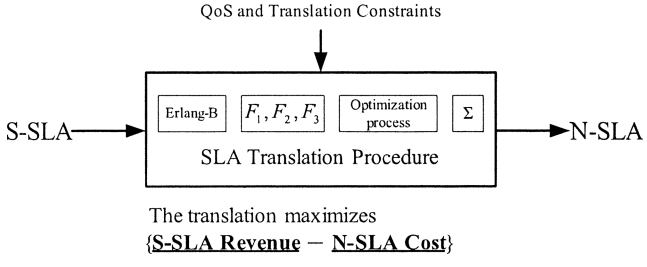


Fig. 4 The reference model of SLA translation.

The procedure of SLA translation is shown in Fig. 4. We only estimate the network resource requirements of simplex-channel communication here. Because VoIP services are duplex-channel communication, the requirements of the opposite channels can be estimated similarly. Besides, we assume that the traffic engineering of VoIP service are preplanned well. Because the VoIP service provider would contract the S-SLA with many VoIP customers, he should evaluate the VoIP traffic between each pair of edge LSRs (Label Switching Router) and contract the N-SLAs with one or several network provider. Thus, the VoIP traffic of many VoIP customers between a certain pair of edge LSRs contracted in different S-SLAs would be aggregated into one or few LSP according to E-LSP or L-LSP technology [18]. First, in the below description, we will estimate the network resource requirements of each PHB class for a S-SLA between a certain pair of edge LSRs. Finally, the total network resource requirements of each PHB class between each pair of edge LSRs can be estimated.

The procedure of SLA translation between a certain pair of edge LSRs for a S-SLA is explained as below.

1. The amount of lines for the customers which is denoted as L can be estimated by Erlang-B formula according to BHT and SBP. The quantity in the unit of "line" can be decided. In this paper, line is defined as a simplex virtual connection to carry user data. For example, a general call contains two lines with different direction, and a line may carry an audio steam and a video stream. Let n_i denote the number of lines operating in Q_i and it is defined in Eq. (1).

$$n_i \equiv \sum_{j=1}^{L} Q_{ij} \geq 0, \tag{1}$$

where

$$1 \leq i \leq h \text{ and}$$
$$Q_{i,j} = \begin{cases} 1, & \text{if line } j \text{ operating in } Q_i \\ 0, & \text{otherwise.} \end{cases}$$

2. In order to maximize the profit, the optimal quality of each line can be decided by solving the SLA optimization problem. Therefore, this step is the most important process in the SLA translation procedure.
The main concept of SLA translation is illustrated in Fig. 5. We let q_j, r_j, u_j, c_j denote the operating quality, the resource requirement, the session revenue, the resource cost of Line j respectively. Then r_j, u_j and c_j can be derived from F_1, F_2 and F_3 that $r_j = F_1(q_j)$, $u_j = F_2(q_j)$, $c_j = F_3(r_j)$. Thus, we can summarize the

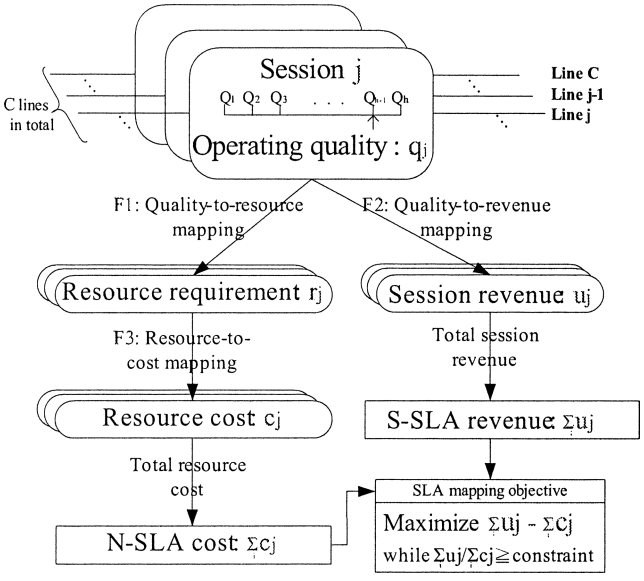


Fig. 5 The main concept of SLA translation.

resource costs (c_j) and session revenues (u_j) to evaluate the N-SLA cost and S-SLA revenue.
The optimization problem can be formulated as Eq. (2) to solve each q_j.

Objective:

$$\text{Maximize} \sum_{j=1}^{L} u_j - \sum_{j=1}^{L} c_j \tag{2}$$

Subject to:

$$\frac{\sum_{j=1}^{L} u_j}{\sum_{j=1}^{L} c_j} \geq \text{minimum revenue-to-cost ratio} \tag{3}$$

$$n_i = \sum_{j=1}^{L} Q_{ij} \geq N_i \tag{4}$$

where

$$\sum_{j=1}^{L} u_j = \sum_{j=1}^{L} F_2(q_j)$$

$$\sum_{j=1}^{L} c_j = \sum_{j=1}^{L} F_3(r_j) = \sum_{j=1}^{L} F_3(F_1(q_j))$$

3. After deciding the operating quality of each line (q_j), the network resource requirements can be evaluated according to F_1. The resource requirement (R_k) of each PHB class (k) can be calculated as

$$R_k = \sum_{j=1}^{L} \hat{R}_{jk}, \tag{5}$$

where

$$1 \leq k \leq n \text{ and}$$

$$\hat{R}_{jk} = \begin{cases} r_j, \text{ if line } j \text{ is operating in PHB class } k \\ 0, \text{ otherwise.} \end{cases}$$

4. The R_ks of other S-SLAs could be evaluated similarly, and the N-SLA between the pair of edge LSRs would be contracted thus.

```
algorithm SLAs_Translation();
begin
    Initialize a solution that satisfies Ni for 1 ≤ i ≤ h
    and has the smallest profit in other (L − ∑Ni) lines;
  for line index from 1 to (L − ∑Ni) do
      for quality index from initial solution +1 to
      the number of the quality levels do
          if current solution +1 satisfies mapping
          constraints and it has the maximum profit
          then update the solution with the maximum
               profit;
          else continue with next line;
      repeat
  repeat
end
```

Fig. 6 The heuristic algorithm for S-SLA and N-SLA translation.

6.1.2 Heuristic Algorithm

The optimization problem described in the previous section can be resolved by finding the number of lines operating in each service level (n_i) that maximizes the profit while satisfying the QoS and translation constraints. The optimal formulation is a non-linear Integer problem and cannot be resolved directly. With the exhaustive search method, all feasible profits with every line operating on each quality profile have to calculate, and then the optimal solution can be decided. However, the complexity of the exhaustive search method is $O(h^L)$, where h denotes the amount of quality profiles and L denotes the total amount of lines. The optimization problem is a NP (nondeterministic polynomial time) problem. Accordingly, the heuristic algorithm designed for SLA translation is presented in Fig. 6, and the complexity can be reduced to $O(h \times (L - \sum N_i))$.

First, we fix the number of lines operating in each Q_i according to N_i in S-SLA and initialize the profit to the minimum by assigning the remain $(L - \sum N_i)$ lines to operate in the lowest operating quality, i.e., Q_1. Second, the initial operating qualities of $(L - \sum N_i)$ lines are gradually promoted to the proper operating qualities that produce more profit. For each line index from 1 to $(L - \sum N_i)$, it upgrades the quality level from initial solution to the highest quality level. The decision criterion of upgrading the quality level is to maximize the profit while satisfying mapping constraints. After iterating all the $(L - \sum N_i)$ lines, the heuristic procedure is done. An example is examined in the next section for more explanation.

Table 1 A S-SLA example.

Item	Content
Busy Hour Traffic (BHT)	3.1 Erlang
Session Blocking Probability (SBP)	0.1%
Callee parties	communicate with a certain enterprise
Quality Profile (Q_i)	Q_1,Q_2 and Q_3 are provided
Guaranteed Quantities (N_i)	$N_1 = 1$, $N_2 = 2$ and $N_3 = 3$
User Options	none
Billing Mechanism	the costs of Q_1, Q_2 and Q_3 are equal to 0.03, 0.075 and 0.13 yens/second respectively
Regular Fee	0
Discounting policy	0
Others	none

6.1.3 Numerical Example

A numerical translation example of one S-SLA and one N-SLA is described as below.

1. We assume the S-SLA is contracted as Table 1 in advance. The call arrival rate is equal to 0.0155 call/second, the average call holding time is equal to 200 seconds, and then Erlang value is equal to 3.1. Thus, according to Erlang-B formula, the total amount of lines L is equal to 10 lines.
2. The pricing strategies (F_2) and the resource allocation scheme (F_1) for VoIP service are actually up to the VoIP provider, whereas the pricing strategies for network service (F_3) depend on the network service provider. However, the exploration of F_1, F_2 and F_3 falls out the interest of this paper. The quality-to-revenue mapping (F_2) presented here is based on the expected revenue generated by each session operating quality, for example, the expected revenues of Q_1, Q_2 and Q_3 are equal to $0.03 \times 200 = 6$, $0.075 \times 200 = 15$, and $0.13 \times 200 = 26$ yens respectively. Additionally, it has the tendency that the better quality a VoIP session operates in, the more resource and the higher transmission quality the VoIP session requires, i.e., Q_3 operating in a better video/audio quality inherently requires more bandwidth and higher transmission quality than Q_1 and Q_2.
   Additionally, we assume 128 kbps in BE, 256 kbps in AF, 384 kbps in EF can satisfy the requirements of Q_1, Q_2 and Q_3 respectively. Therefore, the network cost incurred by Q_1, Q_2 and Q_3 are equal to 1.5, 5 and 13 respectively.
   Besides, we also assume that the minimum revenue-to-cost ratio desired by the VoIP service provider is equal to 2.3. To find optimal operating quality of each line, we use the heuristic optimization algorithm specified in the previous section. Each iterative state of the SLA translation is shown in Table 2. Finally, $(1, 5, 4)$ is the optimal solution, and the amounts of each operating quality are $n_1 = 1$, $n_2 = 5$ and $n_3 = 4$ respectively. The profit is equal to 106.5.
3. After getting amounts of each operating quality, the resource requirement of each PHB class can be calculated by Eq. (5). The resource requirements of EF class, AF class and BE class are $4 \times 384 = 1536$ kbps, $5 \times 128 = 640$ kbps and 128 kbps respectively.
4. The total amount of resource required (R) can be calculated as $R = \sum_{k=1}^{3} R_k = 2304$ kbps.

Table 2 The iterative states of SLA translation using heuristic algorithm.

Step	(n_1, n_2, n_3)	Profit	Revenue-to-Cost Ratio	Meet Constraint
Initial	$(5, 2, 3)$	81.5	2.44	Acceptable
Line 1	$(4, 3, 3)$	87	2.45	Acceptable
	$(4, 2, 4)$	90	2.32	Acceptable
Line 2	$(3, 3, 4)$	95.5	2.33	Acceptable
	$(3, 2, 5)$	98.5	2.23	Fail
Line 3	$(2, 4, 4)$	101	2.34	Acceptable
	$(2, 3, 5)$	104	2.25	Fail
Line 4	$(1, 5, 4)$	106.5	2.35	Acceptable
	$(1, 4, 5)$	109.5	2.26	Fail

6.2 On-Line Adaptive QoS Tuning

After contracting proper N-SLA in the previous sections, on-line adaptive QoS tuning design challenges to a complex problem of assigning proper operation quality dynamically to each session. Therefore, the goal of on-line QoS tuning is to adapt the VoIP service policy to the system condition changes, e.g., the changes of network loading and system loading. In a realistic environment, F_1' is changed with network loading any time and estimated periodically, but F_2' and F_3' are usually stable and equal to F_2 and F_3 in the off-line evaluation.

In the paper, we propose two objectives to on-line adaptive QoS tuning. The first is to maximize the profit as well as off-line QoS evaluation, and the other is to minimize the penalty of wasting network resource [21] or violating S-SLA commitments. If the resource of each PHB class can be controlled well, the most users can receive the proper quality, the resource utilization can be improved and the system capacity can be increased.

Our service principle in on-line process is to satisfy the user's requirements first such as guaranteed quantities (N_i), and then to maximize the profit of VoIP provider. The operating quality of the out-profile call can be upgraded to improve the resource utilization at low system loading, and it also can be degraded to increase the system capacity at high system loading since network resource has been reserved in the N-SLA. When a new call j arrives, the operating quality of the call will be decided by VoIP users if the contracted quota of the selected operating quality is available. Otherwise, the proper operating service of the call will be dynamically and compulsorily decided during the call setup by the VoIP call server. If the objective of on-line process is to maximize the profit of VoIP provider, each session in arrival or in conversation can be tuned to a proper operation quality

dynamically according to our service principle.

To deal with time-critical problem of on-line process, the computation can not be triggered by call arrival. The on-line SLA optimization problem is background pre-computed with extra m sessions, where $m = 1$ to M. If the burst arrivals are less than m sessions, all of them can be handled. As time goes on, the number of computations will be decreased since policy database is filled up with various system conditions.

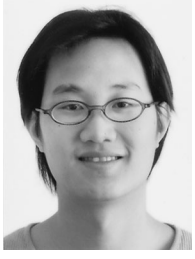
7. Conclusion

This paper proposes a framework of session-level SLA and network-level SLA management to provide QoS-oriented application services over DiffServ/MPLS networks. The SLA management framework contains off-line evaluation and on-line process. In the phase of off-line evaluation, we propose the templates of S-SLA and N-SLA structures and the SLA translation mechanism to provide a N-SLA contracting guideline to a VoIP provider. Additionally, the heuristic algorithm is presented to resolve the optimization problem efficiently. In the phase of on-line process, we introduce the service policy of assigning the proper operation quality to each call. Briefly, the robust, fair, flexible and efficient SLA management framework is produced to application service providers in the paper. Additionally, the general management framework can be approached in other real-time multimedia and non-real time data services, and the problem of on-line SLA process can be resolve with an simple optimization method to improve its performance efficiently in the future.

References

[1] T.M.T. Nguyen, N. Boukhatem, Y. Doudane, and G. Pujolle, "COPS-SLS: A service level negotiation protocol for the Internet," IEEE Commun. Mag., vol.40, no.5, pp.158–165, May 2002.
[2] T.M.T. Nguyen, N. Boukhatem, and G. Puiolle, "COPS-SLS usage for dynamic policy-based QoS management over heterogeneous IP networks," IEEE Netw., vol.17, no.3, pp.44–50, May–June 2003.
[3] J. Skene, D.D. Lamanna, and W. Emmerich, "Precise service level agreements," Proc. IEEE 26th International Conference on Software Engineering, pp.179–188, May 2004.
[4] D. Verma, "Service level agreements on IP networks," Proc. IEEE, vol.92, no.9, pp.1382–1388, Sept. 2004.
[5] E. Bouillet, D. Mitra, and K. Ramakrishnan, "The structure and management of service level agreements in networks," IEEE J. Sel. Areas Commun., vol.20, no.4, pp.691–699, May 2002.
[6] J.-T. Park, J.-W. Baek, and J.W.K. Hong, "Management of service level agreements for multimedia internet service using a utility model," IEEE Commun. Mag., vol.39, no.5, pp.100–106, May 2001.
[7] E. Marilly, O. Martinot, S. Betge-Brezetz, and G. Delegue, "Requirements for service level agreement management," Proc. IEEE Workshop on IP Operations and Management, 2002.
[8] TM Forum, "SLA management handbook," GB917 v2.0, Tech. Rep., Feb. 2004.
[9] E. Marilly, O. Martinot, S. Betge-Brezetz, and G. Delegue, "Requirements for service level agreement management," Proc. IEEE Workshop on IP Operations and Management, pp.57–62, 2002.
[10] L. Lewis and P. Ray, "Service level management definition, architecture, and research challenges," Proc. IEEE GLOBECOM, pp.1974–1978, 1999.
[11] H. Furuya, S. Nomoto, H. Yamada, N. Fukumoto, and F. Sugaya, "Toward QoS management of VoIP: Experimental investigation of the relations between IP network performances and VoIP speech quality," IEICE Trans. Commun., vol.E87-B, no.6, pp.1610–1622, June 2004.
[12] L. Lundy and R. Pradeep, "On the migration from enterprise management to integrated service level management," IEEE Netw., vol.16, no.1, pp.8–14, Jan.–Feb. 2002.
[13] I. Yamasaki, R. Kawamura, and K. Iwashita, "A profit maximization scheme by service-list control for multiple class services," IEICE Trans. Commun., vol.E87-B, no.5, pp.1334–1345, May 2004.
[14] H.K. Su, C.S. Wu, and K.J. Chen, "Session classification for traffic aggregation," Proc. IEEE ICC, pp.1243–1247, June 2004.
[15] C. Molina-Jimenez, S. Shrivastava, J. Crowcroft, and P. Gevros, "On the monitoring of contractual service level agreements," Proc. IEEE Workshop on Electronic Contracting, pp.1–8, July 2004.
[16] TM Forum, "Performance reporting concepts & definitions document," TMF701, v2.0, Tech. Rep., Nov. 2001.
[17] A. Takahashi, H. Yoshino, and N. Kitawaki, "Perceptual QoS assessment technologies for VoIP," IEEE Commun. Mag., vol.42, no.7, pp.28–34, July 2004.
[18] F.L. Faucheur, B.D.L. Wu, S. Davari, P. Vaananen, R. Krishnan, P. Cheval, and J. Heinanen, "Multi-protocol label switching (MPLS) support of differentiated services," RFC 3270, May 2002.
[19] B. Davie, A. Charny, J. Bennet, K. Benson, J.L. Boudec, W. Courtney, S. Davari, V. Firoiu, and D. Stiliadis, "An expedited forwarding PHB (Per-Hop Behavior)," RFC 3246, March 2002.
[20] J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski, "Assured forwarding PHB group," RFC 2597, June 1999.
[21] H.K. Su, H. Chen, C.Y. Wang, and K.J. Chen, "A novel AF+ service for VoIP applications over a Diff-Serv/MPLS network," Proc. IEEE VTC 2004-Fall, pp.4846–4850, Sept. 2004.

Hui-Kai Su was born in Chia-Yi, Taiwan, in 1977. He received the B.S. degree from I-Shou University, Taiwan, in 1999. He received the M.S. degree and still works on his Ph.D. degree at National Chung-Cheng University. His research interests include QoS control, resource management and survivability about IP/MPLS networks and multimedia applications.



Zhi-Zhen Yau was born in Taipei, Taiwan, in 1980. He received the B.S. degree from National Chung-Cheng University, Taiwan, in 2003. He is still working on his M.S. degree at National Chung-Cheng University. His research interests include VoIP, SLA management and IP/MPLS networks.

Cheng-Shong Wu was born in Taoyuan, Taiwan, in 1961. He received the B.S. and M.S. degrees in electrical engineering in 1983 and 1985 from National Taiwan University, Taiwan, and Ph.D. degree in electrical engineering from University of Southern California, USA, in 1990. During 1991, he was a Visiting Assistant Professor in the Center for Advanced Computer Studies in the University of Southwestern Louisiana. He joined the Department of Electrical Engineering at National Chung-Cheng University, Taiwan, in fall of 1991. Currently he is a Professor in the department. His researching interests include broadband networks, wireless networks, queuing theory, and mathematical programming.

Kim-Joan Chen received the B.S. degree in mathematics from National Central University, Taiwan in 1977, and M.S. and Ph.D. degrees in applied mathematics from the State University of New York at Stony Brook in 1981 and 1983, respectively. He was an assistant professor at the Department of Mathematical Science, University of Cincinnati, Cincinnati, Ohio in 1983. During 1984 to 1985, he was with UNINET/TELENET research, and he was with AT&T Bell Laboratories from 1986 to 1989. Since 1989 he has been with the National Chung Cheng University. He served as the Chairperson of the department of Electrical Engineering from 1989 to 1991, as the Director of Computer Center from 1994 to 1998, and as the Director of the Computer Center in the Ministry of Education, Taiwan, R.O.C. from 2000 to 2004. His research interests include control and management of computer and telecommunication networks, high-speed networks, and wireless multimedia communications.