Available online at www.sciencedirect.com
ScienceDirect
Computer Communications 30 (2007) 3055–3062
computer communications
www.elsevier.com/locate/comcom

An objective-oriented service model for VoIP overlay networks over DiffServ/MPLS networks

Huan Chen [a], Hui-Kai Su [b,*], Bo-Chao Cheng [a]

[a] Department of Electrical Engineering, National Chung-Cheng University, No. 160 San-Hsing, Min-Hsiung, Chia-Yi 621, Taiwan
[b] Department of Computer Science and Information Engineering, Nanhua University, No. 32, Chung Keng Li, Dalin, Chia-Yi 622, Taiwan

Available online 23 June 2007

Abstract

Bandwidth provisioning and QoS mapping are key issues to support multimedia services such as VoIP in the emerging network technologies. The classification feature enables MPLS (Multi-protocol Label Switch) to support differentiated types of services (DiffServ) with needed QoS. The differentiated service model provides a variety of mechanisms to achieve different objectives (such as call level or packet level satisfaction). To design a good service model, which can balance the call level and packet level QoS performances, is a challenging task. In this paper, we propose two objective-oriented service models for the VoIP services over the DiffServ/MPLS networks. They can be modeled as the *continuous time Markov chains* (CTMC) and the performance are assessed in details. The salient point for the proposed service models is to solve the myth of the trade-off between the service quality (user's concern) and the system revenue (system provider's concern) – involving how to meet each user's SLA requirements while maximizing system revenue. The analytical results in this paper can provide useful information to both user and service provider for signing a cost-effective contract which provides a better trade-off between cost and QoS.

Keywords: Bandwidth provision; QoS mapping; Service model; VoIP; Overlay network; MPLS; DiffServ

1. Introduction

An overlay network is a logical application-layer topology established by overlay nodes. The logical connections between the overlay nodes are provided by overlay links, each of which may be a long path traversing multiple routers and physical links in the Internet. VoIP (Voice over IP) service is one of the peer-to-peer applications of overlay network, and constructs a VoIP service overlay network by VoIP service components and VoIP user agents. The VoIP service overlay network can effectively use the Internet as a lower level transport network to provide VoIP services to end users.

Nowadays, VoIP applications [1] are ubiquitous in the Internet due to its low cost compared to conventional voice service via Public Switched Telephone Network (PSTN). The VoIP service overlay networks have emerged as a means to enhance end-to-end availability and quality of service (QoS). The services require not only adequate bandwidth to support voice transmission but also end-to-end transmission quality between end users. Inherited from its interactive and real-time nature, it is very sensitive to network congestions and it requires more stringent QoS than data traffic. Some QoS requirements shall be maintained to make them attractive to users [2], for example, good voice quality and short call waiting time.

Internet and network communities have addressed the QoS issues about real-time applications for a long time. The emergence technology known as "Differentiated Service Model over Multi-protocol Label Switch" (DiffServ/MPLS) [3] can enhance QoS provisioning ability for the conventional IP-based networks. The DiffServ-based

* Corresponding author. Tel.: +886 5 272 1001x50202; fax: +886 5 242 7131.
E-mail addresses: huan@ee.ccu.edu.tw (H. Chen), hksu@mail.nhu.edu.tw (H.-K. Su), bcheng@ccu.edu.tw (B.-C. Cheng).

0140-3664/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.comcom.2007.05.037

MPLS network is a suitable platform to provide VoIP services, since it can support differentiated traffic classes and provide preferential treatments to users. In addition, such network also makes it possible to provide many salient functions such as QoS provisioning, Fast Forwarding, Traffic Engineering (TE) and Virtual Private Network (VPN) applications [4–7]. The DiffServ-based MPLS technology is scalable and practical; therefore, this technology has been deployed in most advanced routers. With QoS provisioning, this technology can realize the QoS guarantee of VoIP service overlay network very well.

A VoIP service provider purchases bandwidth with certain QoS guarantees from individual network domains via Service Level Agreement (SLA) to build a logical VoIP service overlay network on top of existing data transport networks. Via a service contract, the VoIP service provider directly receives the revenue from VoIP users. First, besides service discovery and service interworking, the VoIP service provider faces the issues of QoS guarantee and service profit. One way to accommodate users' resource demanding, thus to guarantee their demand volume and QoS, is to employ new infrastructure to increase extra many bandwidth to avoid congestions; however, such business strategy is not cost-effective. An alternative prominent strategy to meet the QoS requirements is to raise the system utilization of a network by doing the bandwidth provisioning. In addition, network provisioning can give users better QoS, avoid network congestions and maximize the network efficiency.

Second, in DiffServ/MPLS networks, the VoIP service provider also faces the QoS-mapping issue between overlay network and transport network. The conventional VoIP service providers support only either one of the three service classes in a single VoIP trunk to a group of VoIP subscribers, i.e., *Expedited Forwarding* (EF)-class, *Assured Forwarding* (AF)-class, and *Best Effort* (BF)-class. EF-class traffic receives the highest forwarding priority while BF-class traffic receives the lowest forwarding priority among the three. Transport core networks intend to provide EF-class users with low delay, low jitter and low loss services by serving them at a configured rate [8]. In contract, an AF-class may be configurable to receive more bandwidth resource to forward packets only when excess resource is available [9]. Excess resource, if still available, is left to BE users so that they will not be dried-up. However, although the EF-class provides the best QoS, the EF resource costs the most. Similarly, the AF resource cost a little less than that of EF-class and thus provide degraded service; BE is the cheapest one, no QoS shall be assumed and unsuitable for VoIP applications.

Obviously, it is difficult for VoIP users to choose an cost-effective service among them, especially in a time varying network conditions. Consequently, users either pay more to sign a higher priority service when transport network is not congested, or pay less to receive lower priority service and thus get much poorer throughput than expected.

To help users to receive appropriate services, we propose two objective-oriented service models AF^{+} in this paper. They are modeled as the continuous time Markov chains (CTMC), and the analytical results in this paper can provide useful information to both user and service provider for signing a cost-effective contract which provides a better trade-off between cost and QoS.

The remainder of this paper is organized as follows. In Section 2, the related works are briefly reviewed. A VoIP service overlay network architecture is explained in Section 3. In Section 4, we propose new AF^{+} service models: one is for non-rate adaptive (*non-RA*) policy and the other is for rate-adaptive (*RA*) policy. SLA can be contracted with either *non-RA* or *RA* policy. Both service models are proposed in Sections 4.1 and 4.2, respectively. Performance analysis for both models is illustrated in Section 5. Finally, Section 6 concludes this paper.

2. Related works

A lot of Internet standards on VoIP service architecture and interworking can be found in [10–12]. VoIP services still consider the issues of performance, routing path selection, fault detection and QoS between overlay network and transport network to satisfy users' requirements. In an overlay network, data transfer might not be as efficient as the one performed at the network layer. Service routing overhead is a key performance metric for overlay infrastructures. Zhang et al. [13] proposed a novel mechanism, mOverlay, for constructing an overlay network that takes account of locality of network hosts, and Han et al. [14] proposed a novel framework for topology-aware overlay networks. Additionally, the approaches of topology design to improve the performance of overlay networks can be found in [15]. In our work, we assume the underlying DiffServ/MPLS networks can support the abilities of network resource reservation and QoS guarantee well. The VoIP service providers purchase network bandwidth from several network domains to construct a VoIP service overlay network according to the VoIP service contract of VoIP users. However, the topology design of VoIP service overlay network falls out of the interest of this work and it can refer to the above papers.

After constructing the overlay network, another key problem in the overlay network deployment is the issue of QoS management and bandwidth provisioning, which is critical to cost recovery in deploying and operating the VoIP services over the overlay network. The approaches of QoS-aware routing can be found in [16,17]. Additionally, the research on resource management and bandwidth provisioning can be found in [18–21]. When many parties share network resources on an overlay network, mechanisms must exist to allocate the resources and protect the network from overload. Amir et al. [18] demonstrated near-optimal utilization of network resources, fair sharing of individual congested links, and quick adaptation to network changes. Duan et al. [20] studied the bandwidth provisioning prob-

lem for a service overlay network which is critical to the cost recovery in deploying and operating value-added services over the service overlay network, and mathematically formulate the bandwidth provisioning problem, taking into account various factors such as SLA, service QoS, traffic demand distributions, and bandwidth costs.

In DiffServ/MPLS networks, the EXP-Inferred-PSC LSPs (E-LSP) and Label-Only-Inferred-PSC LSPs (L-LSP) solutions [3] have been developed to support the DiffServ service models (such as EF, AF and BE) to enable the MPLS network classifying services of various applications. However, LSPs can only receive the same class for all packets in both E-LSP and L-LSP, which cannot solve the myth of trade-off between cost and service quality. Additionally, the ideal of AF^{+} service model was presented in [22] first. To provide cost-effective DiffServ solutions in MPLS networks, we proposed two service models, $AF^{+}_{non\text{-}RA}$ and AF^{+}_{RA}, to extend the conventional AF service for non-rate adaptive and rate adaptive call admission policies, respectively.

3. Overlay network architecture

A VoIP service overlay network architecture is illustrated in Fig. 1. According to VoIP service contracts, the VoIP providers purchase bandwidth with certain QoS guarantees from underlying network domains to build a logical VoIP service overlay network on top of existing data transport networks. In the transport network, the VoIP users can access broadband networks through various technologies such as xDSL, Cable Modem, Fiber-to-the-Home (FTTH), etc. They can receive the high quality and high speed transport service to attach their VoIP service provided by the local VoIP service provider.

The VoIP service overlay network topology is established by VoIP service components and their users. The logical connections between the service components and the user agents are provided by overlay links, each of which may be a long path traversing multiple routers and physical links in the transport network. Particularly, the VoIP service overlay network can be divided into two sublayers: signaling sublayer and media sublayer. The signaling sublayer network is constructed by several signaling servers, which may be deployed by multiple VoIP service providers. The signaling servers provide the functionalities of registrar server, location server and proxy server in the SIP standard [10]. The media sublayer network is established by media gateways, which may be deployed by different VoIP service providers. The media gateways keep the states of each VoIP sessions, and handle each media stream of VoIP sessions. Moreover, they classify the VoIP packets and aggregate the media streams into their proper QoS-enabled overlay path (i.e., DiffServ-based LSP) [23]. Therefore, if the bandwidth provisioning (quantity) and QoS mapping (quality) is control well, the requirements of each VoIP user can be satisfied in the consideration of resource cost.

For example, when the UA1 is powered on, it attaches to its local SS1 and performs the registration procedure. The signaling servers manage the location information and states of their UAs. If UA1 makes a call to UA2, SS1 will receive the request from UA1 and then help him to discover the location of UA2. The signaling negotiation is delivered via the overlay paths of signaling sublayer with dedicated network resource. After UA2 answers this call, SS1 and SS3 separately handle MG1 and MG4 to aggregate the media streams of this call into the proper routing overlay path with certain quality level in media sublayer. Finally, after UA1 or UA2 releases this call, SS1 and SS3 also separately handle MG1 and MG4 to release the traffic aggregation.

4. System and service models

The conventional VoIP service model supports only either one of the three priority classes in a single VoIP trunk for a group of VoIP subscribers. It is difficult for users to determine which service class to sign up will do themselves the best. Similarly, the service providers do not have the flexibility to negotiate with users for their services. Let's elaborate the above situation by an example.

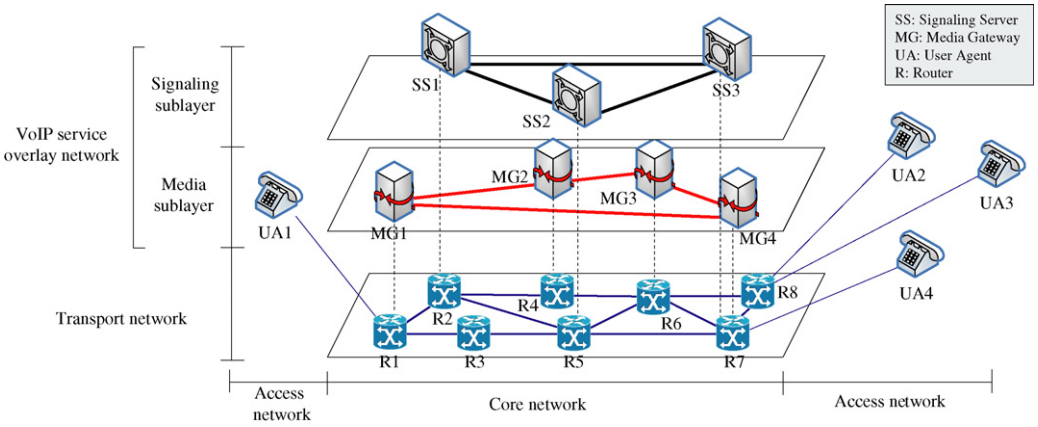

Fig. 1. A VoIP service overlay network architecture.

The VoIP service provider spends high cost and purchases network bandwidth with EF-class service to deploy a VoIP trunk. Consider a VoIP user requests a VoIP service when the VoIP trunk contracted previously is full. Because of the hard guarantee of EF-class service, this call will be blocked. In the contrast, the service provider spends less cost to purchase AF-class bandwidth than EF-class bandwidth. If he VoIP trunk contracted previously is congested and the AF-class bandwidth of other VoIP trunks is available, the new call will be accepted and share the remaining AF-class bandwidth of others. However, the quality of AF-class VoIP trunk is unstable and dependent on other network users. The conventional service model do not support such functionality, which motivates us to propose the following generic AF^+ service model.

Here, we first propose a generic AF^+ service model which is constructed based on the DiffServ/MPLS networks but with an adaptive SLA signing policy. VoIP providers purchase bandwidth with EF-class service and AF-class service simultaneously to provide AF^+ VoIP service. AF^+ provides *EF*-class when VoIP trunk is not congested and it provides only *AF*-class service when VoIP trunk is congested. Consider the following scenario, for example, when a new call arrives and the provisioned bandwidth of EF-class is available, this call is served with the EF-class level first (up to m *EF* calls). Otherwise, if the provisioning bandwidth of EF-class is unavailable, it is serviced with the AF-class level (up to n *AF* calls) and then is treated as an AF-class like that in the conventional DiffServ/MPLS model; when VoIP trunk is congested, an *EF* request will be degraded to an *AF*-class with discount charge. The design goal is to reduce the bandwidth demanding from a single user when VoIP trunk is congested, and thus increase the availability of services. As an example illustrated in Fig. 2, there are two VoIP trunks (i.e., overlay links or paths of media sublayer) in a link. In a VoIP trunk, the first coming VoIP streams receive *EF* service up to m streams and the rest of n streams receive only *AF*-class service. Additionally, the extra streams may share the remaining AF-class bandwidth with other users in the link. Assume that the total bandwidth in the VoIP trunk is denoted as BW_{total} which can be expressed as the following.

$$BW_{total} = (m \cdot BW_{EF} + n \cdot BW_{AF}) \tag{1}$$

where BW_{EF} and BW_{AF} are the bandwidth allocated to each EF and AF-class user, respectively. In this example, the ratio of serving *EF* traffic in a single trunk is controlled by a parameter of bandwidth provisioning ratio, α, where α is the VoIP preferential treatment threshold and can be expressed as

$$\alpha = \frac{m \cdot BW_{EF}}{BW_{total}} \quad 0 \leqslant \alpha \leqslant 1 \tag{2}$$

In other words, if a total amount of bandwidth for a VoIP trunk is BW_{total}, the total amount of bandwidth used for serving *EF* traffic is $\alpha \cdot BW_{total}$. An appropriate value of α can be estimated based on the traffic loading and the cost for each class. The effect on α will be discussed in Section 5. In the following paragraphs, we introduce and analyze its two variants, denoted as $AF^+_{non\text{-}RA}$ and AF^+_{RA}, respectively. They differs in their SLA policies. $AF^+_{non\text{-}RA}$ is an AF^+ service model without rate-adaptive feature for calls. AF^+_{RA} is the rate-adaptive version for AF^+ model, in which the *AF* service could be upgraded to *EF* service when the VoIP trunk utilization drops down below α.

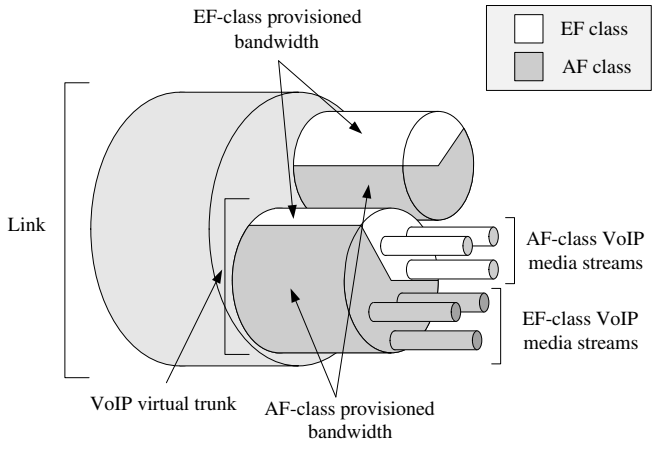

Fig. 2. EF-class and AF-class provisioned bandwidth in a VoIP trunk and link.

4.1. $AF^+_{non\text{-}RA}$: non-rate adaptive service model

$AF^+_{non\text{-}RA}$ follows the same differentiated service framework as AF^+ service model, i.e., the first m calls will be serviced with *EF*-class, and after that all service requests will be treated as *AF*-class up to n calls. But in this $AF^+_{non\text{-}RA}$ model, the *AF*-class service, once determined by DiffServ ingress router while call setup progress, will never be upgraded back to *EF*-class even the system utilization drops back to the ratio of α. In this service model, the ingress routers trace VoIP control messages, manage the states of VoIP sessions and decide which service class they are before VoIP connection is established.

The proposed $AF^+_{non\text{-}RA}$ service can be modeled as a two-dimension birth–death process [24]. Fig. 3 shows the state transition diagram of this birth–death process. We make the following three assumptions before we perform our analysis.

- All traffic patents of VoIP sources are CBR (constant bit rate) with the same coding rate.
- Bandwidth provisioning for m EF-class calls and n AF-class calls are contracted by a VoIP service provider and a network provider in advance.
- Call arrivals for both *EF* and *AF*-classes follow the Poisson distribution with the rate of λ. The call holding time for both *EF* and *AF* services are exponentially distributed with the mean call holding time of $1/\mu$. The *Erlang* density is defined as a traffic load of VoIP calls, in which $Erlang = \lambda/\mu$.

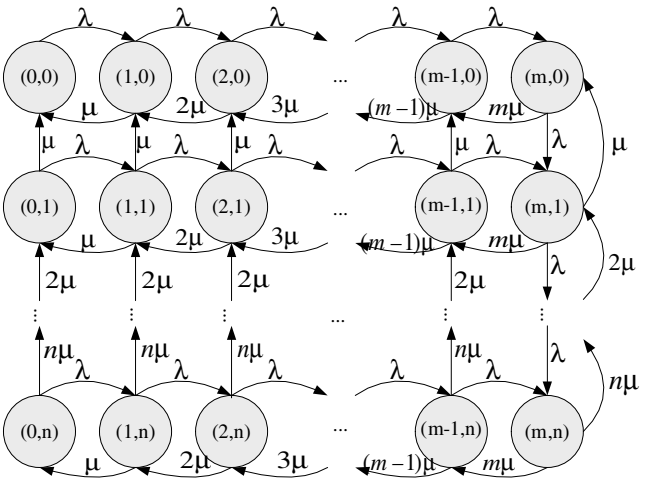

Fig. 3. Two-dimension state-transition-rate diagram for the $AF^{+}_{non\text{-}RA}$ service model.

Observe that in Fig. 3, if a VoIP trunk contains i EF-class and j AF-class, we say that this system is in the state of (i,j), where i and j are positive integers in the ranges of $0 \leqslant i \leqslant m$ and $0 \leqslant j \leqslant n$, respectively. Let $P_{i,j}$ be the stationary sate probability of (i,j), it can be found by solving the equilibrium equations as below.

$$\begin{cases} \lambda P_{0,0} = \mu P_{0,1} + \mu P_{1,0}, & \text{if } i = 0 \text{ and } j = 0 \\ (\lambda + i\mu)P_{i,0} = \lambda P_{i-1,0} + (i+1)\mu P_{i+1,0} + \mu P_{i,1}, & \text{if } 1 \leqslant i \leqslant m-1 \text{ and } j = 0 \\ (\lambda + m\mu)P_{m,0} = \lambda P_{m-1,0} + \mu P_{m,1}, & \text{if } i = m \text{ and } j = 0 \\ (\lambda + j\mu)P_{0,j} = (j+1)\mu P_{0,j+1} + \mu P_{1,j}, & \text{if } i = 0 \text{ and } 1 \leqslant j \leqslant n \\ (\lambda + i\mu + j\mu)P_{i,j} = (i+1)\mu P_{i+1,j} + (j+1)\mu \cdot P(i,j+1) + \lambda P_{i-1,j}, & \text{if } 1 \leqslant i \leqslant m-1 \text{ and } 1 \leqslant j \leqslant n \\ (\lambda + m\mu + j\mu)P_{m,j} = \lambda P_{m,j-1} + \lambda P_{m-1,j} + (j+1)\mu P_{m,j+1}, & \text{if } i = m \text{ and } 1 \leqslant j \leqslant n \end{cases} \tag{3}$$

The normalization equation for the above equilibrium equations is

$$\sum_{j=0}^{n}\sum_{i=0}^{m} P_{i,j} = 1 \tag{4}$$

In the provisioned MPLS network, if the resource is not enough for an AF-class VoIP call, this call will be blocked in the ingress router of the VoIP trunk. The call blocking rate $P^{non\text{-}RA}_{blk}$ can be derived via Eq. (5).

$$P^{non\text{-}RA}_{blk} = P_{m,n} \tag{5}$$

The provisioned bandwidth utilization for the EF-class and AF-class can be expressed as below.

$$U^{non\text{-}RA}_{EF} = \frac{1}{m}\sum_{j=0}^{n}\sum_{i=0}^{m} iP_{i,j} \tag{6}$$

$$U^{non\text{-}RA}_{AF} = \frac{1}{n}\sum_{j=0}^{n}\sum_{i=0}^{m} jP_{i,j} \tag{7}$$

Our main objective is to minimize the cost function $J^{non\text{-}RA}$. The cost function is such defined to reflect the penalty of the bandwidth waste. Since the service admission control is implemented at the call level, and the same fixed amount of bandwidth is reserved for each service packet flow in order to provide a certain level of quality of service. If excess bandwidth is over-reserved, the utilization at the packet level will decrease. The more bandwidth waste the fewer calls can be further admitted.

Let C_{EF} and C_{AF} denote the penalty costs for EF-class and AF-class, respectively, and the normalized weighting factor β is defined as $\beta = \frac{C_{EF}}{C_{EF}+C_{AF}}$. The cost function of *non-RA* service model, $J^{non\text{-}RA}$, can thus be defined as:

$$\begin{aligned} J^{non\text{-}RA} &= \beta(1 - U^{non-RA}_{EF}) + (1-\beta)(1 - U^{non-RA}_{AF}) \\ &= 1 - \beta U^{non\text{-}RA}_{EF} - (1-\beta)U^{non\text{-}RA}_{AF} \end{aligned} \tag{8}$$

4.2. AF^{+}_{RA}: rate adaptive service model

AF^{+}_{RA} is the rate-adaptive version for AF^{+} model, in which the AF service could be upgraded to EF service when the trunk utilization drops down below α. Therefore, the ingress routers have to trace VoIP control messages and manage the states of VoIP sessions anytime. If there is available EF-class bandwidth, the part of AF-class VoIP sessions in the same trunk will be upgraded to EF service. Thus, the system loading of AF^{+}_{RA} service model in the ingress routers would be higher than the other.

The proposed AF^{+}_{RA} service can be modeled as a one-dimension birth–death process [24]. Fig. 4 shows the state transition diagram of this birth–death process. We make the same assumptions as those for AF^{+}_{RA} service model. Call arrivals for both EF and AF-classes still follow the Poisson distribution with the rate of λ. The call holding time for both EF and AF services are exponentially distributed with the same mean call holding time of $1/\mu$. The *Erlang* density is defined as a traffic load of VoIP calls, in which $Erlang = \lambda/\mu$.

Let P_k be the stationary sate probability of k, it can be found by solving the equilibrium equations as below.

$$P_k = \begin{cases} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} P_0, & 0 \leqslant k \leqslant m+n \\ 0, & k > m+n \end{cases} \tag{9}$$

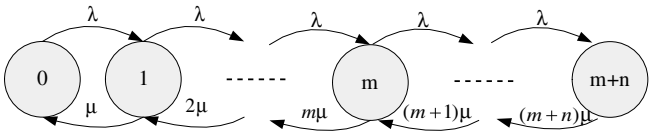

Fig. 4. Two-dimension state-transition-rate diagram for the AF^{+}_{RA} service model.

where

$$P_0 = \left[\sum_{k=0}^{m+n}\left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}\right]^{-1}$$

The normalization equation for the above equilibrium equations is

$$\sum_{k=0}^{m+n} P_k = 1 \tag{10}$$

In the provisioned MPLS network, if the resource is not enough for an *AF*-class VoIP call, this will be blocked in the ingress router of the VoIP trunk. The call blocking rate P_{blk}^{RA} can be derived via Eq. (11) that is alike to Erlang-B Model.

$$P_{blk}^{RA} = P_{m+n} \tag{11}$$

The provisioned bandwidth utilization for the EF-class and AF-class can be expressed as below.

$$U_{EF}^{RA} = \frac{1}{m}\left(\sum_{k=0}^{m} kP_k + \sum_{k=m+1}^{m+n} mP_k\right) \tag{12}$$

$$U_{AF}^{RA} = \frac{1}{n}\sum_{k=m+1}^{m+n}(k-m)P_k \tag{13}$$

The cost function of *RA* service model, AF_{RA}^{+}, similar to which of the *non-RA* service model, is defined as below.

$$\begin{aligned} J^{RA} &= \beta\left(1 - U_{EF}^{RA}\right) + (1-\beta)\left(1 - U_{AF}^{RA}\right) \\ &= 1 - \beta U_{EF}^{RA} - (1-\beta)U_{AF}^{RA} \end{aligned} \tag{14}$$

5. Numerical results

The non-rate adaptive service $AF_{non\text{-}RA}^{+}$ can be modeled as a two-dimensional continuous time Markov chain (2D-CTMC) and the steady-state probabilities of this model can be derived by a sophisticated commercial optimization tool, LINGO. On the other hand, the rate adaptive service AF_{RA}^{+} can be modeled as a one-dimensional continuous time Markov chain (1D-CTMC) and the steady-state probabilities can be easily found by the Erlang-B formula. This section presents the numerical results for both service models. We evaluate the performance of the proposed service model in three aspects: call level, packet level, and an objective function.

5.1. Call level performance analysis

The performance metrics in terms of the call blocking rates and the bandwidth utilization for each model are analyzed. The former represents the service performance at the call level, while the later evaluates the service performance at the packet level.

The call blocking rates are used as the call level performance metric. The traffic intensity of VoIP calls are set to 25 ($Erlang = 25$), and call arrivals are assumed to follow the Poisson distribution. The call blocking rates of $AF_{non\text{-}RA}^{+}$ and AF_{RA}^{+} service models can be derived from Eqs. (5) and (11), respectively. The numerical results are shown in Fig. 5 where the X axis denotes the total provisioned bandwidth ($m+n$); the Y axis stands for the call blocking rates; and the Z axis represents the VoIP preferential treatment threshold (α) defined in Eq. (2).

In this figure, we observe that the blocking rates of call requests depend only on the total bandwidth (totally can accommodate m EF-class and n AF-class calls) regardless of the threshold α, which controls the quality of service at the packet level instead of the dropping probability at the call level. Proposed service models are such designed to maintain the same call level dropping probability for all call requests but provide preferential treatment at packet level. This design principle is important to both customers as well as the system operator of the VoIP applications. From the perspective of the customer, he or she would tolerate degraded services (due to packet dropping) rather than totally been rejected of the call requests. On the other hand, from the perspective of the system operator, it needs to maintain an acceptable level of successful call attempts in order to create ongoing relationships, heightened credibility and repeat business.

5.2. Packet level performance analysis

In this analysis, we assume that the total capacity can accommodate 40 simultaneous call services ($m+n=40$) for both *non-RA* and *RA* service models.

For the *non-RA* model, the total bandwidth utilizations for EF-class and AF-class can be derived from (6) and (7), respectively. Similarly, for the *RA* models, they can be derived from (12) and (13), respectively.

The relationships of bandwidth utilization and VoIP preferential treatment threshold α for two classes are illustrated in Fig. 6. We observe the following two facts: *First*, if a small value of α ($\alpha \to 0$) is selected, the EF-class bandwidth utilization for both models are greater than 0.96. Otherwise, the bandwidth utilization for both of them are

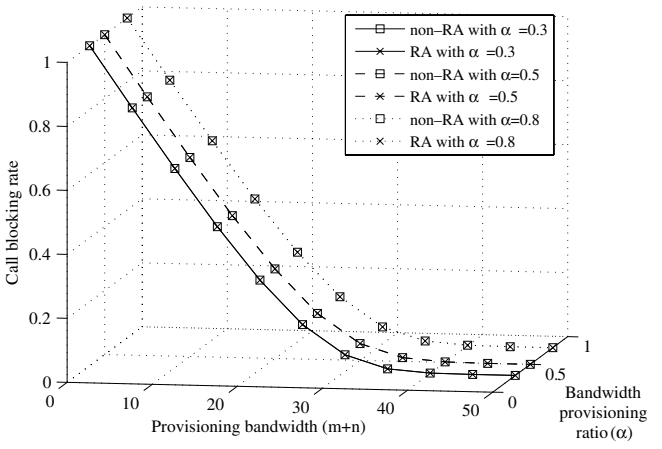

Fig. 5. The call blocking rates of $AF_{non\text{-}RA}^{+}$ and AF_{RA}^{+} service models while $Erlang = 25$.

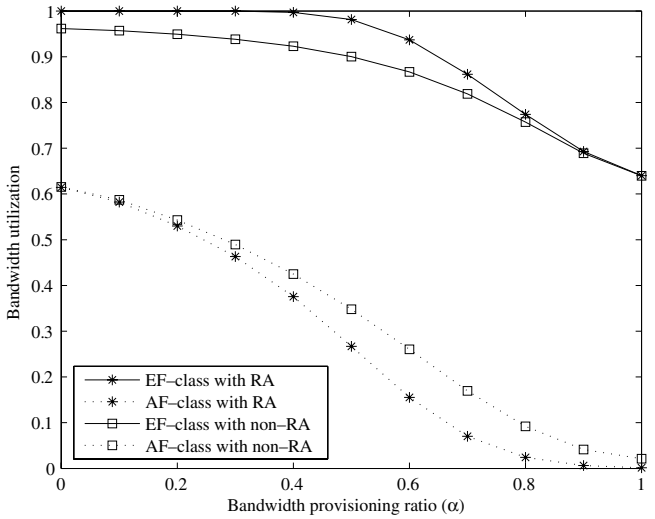

Fig. 6. The comparison of provisioned bandwidth utilizations with $AF^{+}_{non\text{-}RA}$ and AF^{+}_{RA} service models while $m + n = 40$.

decreasing (the larger value of the α, the smaller value of the utilization). It is because the bandwidth allocated to the EF-class is enough to serve most of the call requests. *Second*, the bandwidth utilization for EF-class using *RA* service model is lager than that using the *non-RA* service model. The results reflects the fact because a AF service could be upgraded to a EF service in the *RA* model as long as the bandwidth allocated to the EF-class is available.

5.3. Objective function

The evaluation of the penalty cost function (the objective function) is illustrated in Fig. 7. For both *non-RA* and *RA* services, the cost increases monotonically as the α increases. For a fixed α, *RA* model with the largest value of β performs the best (has the lowest cost). On the other hand, the *non-RA* model with the smallest value of β performs the worst (has the highest cost). In theory, the smaller the α is, the less penalty costs. However, we may need a larger value of α to serve more EF-class service in practice. We observe that the penalty cost increases sharply once the value of α exceeds a "curve point" (an experimental value is between 0.4 and 0.6). We may suggest that the α shall be set to this "curve point" in order to select an appropriate value of α (the VoIP preferential treatment threshold) to optimize the trade-off between the performances at the call level and packet level.

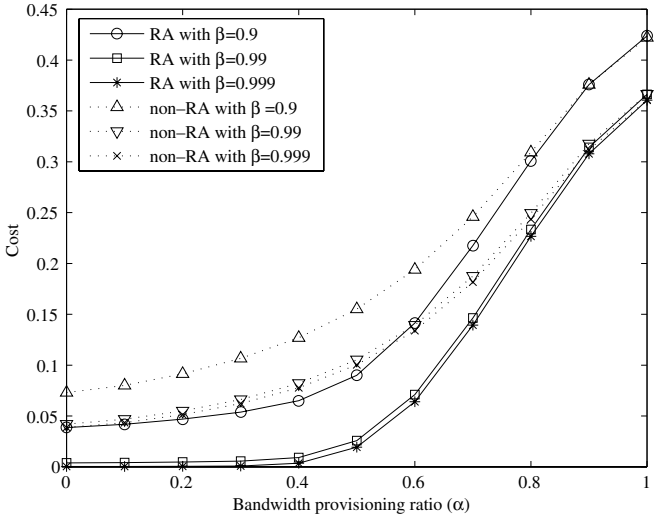

Fig. 7. The comparison of costs with $AF^{+}_{non\text{-}RA}$ and AF^{+}_{RA} service models while $m + n = 40$.

6. Conclusion

Trade-off between the cost and service quality is a well-known SLA problem – it involves how to meet each user's SLA requirements while maximizing system revenue. In this paper, we propose $AF^{+}_{non\text{-}RA}$ and AF^{+}_{RA} service models for VoIP service overlay networks to accommodate emerging applications such as VoIP and real-time video streaming over the DiffServ/MPLS networks. In both service models, incoming calls are first serviced with the EF-class up to a certain threshold (determined by the network service provider) and after that calls would be served with AF-class. For $AF^{+}_{non\text{-}RA}$ service model, each AF call is served with the same class during its call holding time; while for AF^{+}_{RA} service model, an ongoing AF call can be upgraded to an EF call as long as the resource allocated to the entire EF-class is not occupied fully.

The $AF^{+}_{non\text{-}RA}$ and AF^{+}_{RA} services are modeled as the two-dimension and one-dimension Markov chains, respectively, to evaluate their performance under different traffic scenarios. Numerical results show that both proposed AF extension service models take advantages of service guard and cost-savings features from EF and AF-class. The total bandwidths can be assigned to tune the alternative issues of call blocking rate and provisioned bandwidth cost. This study has confirmed that $AF^{+}_{non\text{-}RA}$ and AF^{+}_{RA} meet the cost-effective requirements and provide a practical solution to VoIP service overlay networks. In the future works, the numerical results of quality (i.e., jitter, latency, packet loss) will be demonstrated with simulation or implementation. Both service models also can be easy to apply to other peer-to-peer applications of overlay networks over DiffServ/MPLS networks.

Acknowledgements

This research was supported in part by the National Science Council (NSC) in Taiwan under the Grant Nos. NSC-95-2219-E-194-007, NSC-95-2221-E-194-016 and NSC-95-2218-E-343-002.

References

[1] B. Goode, Voice over internet protocol (VoIP), Proceedings of the IEEE 90 (9) (2002) 1495–1517.

[2] V. Fineberg, A practical architecture for implementing end-to-end QoS in an IP network, IEEE Communications Magazine 40 (1) (2002) 122–130.
[3] F.L. Faucheur, B.D.L. Wu, S. Davari, P. Vaananen, R. Krishnan, P. Cheval, J. Heinanen, Multi-protocol label switching (MPLS) support of differentiated services, RFC 3270, 2002.
[4] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss, An architecture for differentiated services, RFC 2475, 1998.
[5] E. Rosen, A. Viswanathan, R. Callon, Multiprotocol label switching architecture, RFC 3031, 2001.
[6] K. Muthukrishnan, A. Malis, A core MPLS IP VPN architecture, RFC 2917, 2000.
[7] F.L. Faucheur, W. Lai, Requirements for support of differentiated services-aware MPLS traffic engineering, RFC 3564, 2003.
[8] B. Davie, A. Charny, J. Bennet, K. Benson, J.L. Boudec, W. Courtney, S. Davari, V. Firoiu, D. Stiliadis, An expedited forwarding PHB (per-hop behavior), RFC 3246, 2002.
[9] J. Heinanen, F. Baker, W. Weiss, J. Wroclawski, Assured forwarding PHB group, RFC 2597, 1999.
[10] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, E. Schooler, SIP: session initiation protocol, RFC 3261, 2002.
[11] J. Rosenberg, H. Schulzrinne, Session initiation protocol (SIP): locating SIP servers, RFC 3263, 2002.
[12] J. Rosenberg, H. Salama, M. Squire, Telephony routing over IP (TRIP), RFC 3219, 2002.
[13] X.Y. Zhang, Q. Zhang, Z. Zhang, G. Song, W. Zhu, A construction of locality-aware overlay network: mOverlay and its performance, IEEE Journal on Selected Areas in Communication 22 (1) (2004) 18–28.
[14] J. Han, D. Watson, F. Jahanian, Topology aware overlay networks, in: Proc. IEEE INFOCOM 2005, 2005, pp. 2554–2565.
[15] S. Vieira, J. Liebeherr, Topology design for service overlay networks with bandwidth guarantees, in: Proc. IEEE IWQOS 2004, 2004, pp. 211–220.
[16] Z. Li, P. Mohapatra, QRON: QoS-aware routing in overlay networks, IEEE Journal on Selected Areas in Communication 22 (1) (2004) 29–40.
[17] T.-Y. Chung, Y.-D. Wang, D2MST: a shared tree construction algorithm for interactive multimedia applications on overlay networks, IEICE Transactions on Communications E88-B (10) (2005) 4023–4029.
[18] Y. Amir, B. Awerbuch, C. Danilov, J. Stanton, A cost-benefit flow control for reliable multicast and unicast in overlay networks, IEEE/ACM Transactions on Networking 13 (5) (2005) 1094–1106.
[19] Y. Huang, S. Bhatti, Decentralized resilient grid resource management overlay networks, in: Proc. IEEE SCC 2004, 2004, pp. 372–379.
[20] Z. Duan, Z.-L. Zhang, Y.T. Hou, Service overlay networks: SLAs, QoS, and bandwidth provisioning, IEEE/ACM Transactions on Networking 11 (6) (2003) 870–883.
[21] X. Gu, K. Nahrstedt, R. Chang, C. Ward, QoS-assured service composition in managed service overlay networks, in: Proc. IEEE 23rd International Conference on Distributed Computing Systems, 2003, pp. 194–201.
[22] H.-K. Su, H. Chen, C.-Y. Wang, K.-J. Chen, A novel AF+ service for VoIP applications over a Diff-Serv/MPLS network, in: Proc. IEEE VTC 2004-Fall, 2004, pp. 4846–4850.
[23] H.-K. Su, C.-S. Wu, K.-J. Chen, Session classification for traffic aggregation, in: Proc. IEEE ICC 2004, 2004, pp. 1243–1247.
[24] L. Kleinrock, Queueing Systems, Theory, vol. 1, John Wiley & Sons, USA, 1972.

Huan Chen received his B.S. and M.S. (Electrical Engineering) degrees from the National Tsing Hua University (NTHU), Hsing Chu, Taiwan, ROC, in 1993 and 1995, respectively. He earned his Ph.D. (Electrical Engineering) degree from the University of Southern California (USC), Los Angeles, in 2002. He also served as the TA and RA in Signal and Image Processing Institute, Integrated Media Systems Center and Electrical Engineering-Systems Department at University of Southern California during his Ph.D. study. He is an member of IEEE since 1999 and he is author of the book "Radio Resource Management for Multimedia QoS Support in Wireless Cellular Networks", Kluwer Academic Publishers, October 2003 (ISBN: 1-4020-7623-1). In additional to his academic career experiences, he also worked for Xitec Inc., Summitec Co. and Intervideo Inc. as a summer intern and software engineer from June 1997 to January 2001. Since February 2002, he has been with the Electrical Engineering Department at National Chung Cheng University, Chiayi (Taiwan). His principle research interests include QoS support, Ad Hoc routing and network optimization. Other research interests include Ubiquitous computing and network security.

Hui-Kai Su received the B.S. (Electronic Engineering) degree from I-Shou University, Taiwan, in 1999. He earned his M.S. and Ph.D. (Electrical Engineering) degrees from National Chung-Cheng University, Taiwan, in 2001 and 2006, respectively. He is an Assistant Professor of Department of Computer Science and Information Engineering at Nanhua University, Taiwan. His research interests include QoS control, resource management and survivability for multimedia applications in QoS-enabled IP/MPLS networks.

Bo-Chao Cheng is an Assistant Professor of Department of Communication Engineering at National Chung-Cheng University. Cheng received a Ph.D. degree in CIS from New Jersey Institute of Technology in 1996. After graduations, he also worked for Transtech Network (2000–2002), Bellcore (1998–2000) and Racal DataCom (1996–1998), respectively. His broad interests include network security, network management and real-time embedded system design.