Performance Analysis of Bandwidth Provisioning for AF$^+$ VoIP Service Models over DiffServ-based MPLS Networks

Hui-Kai Su, Huan Chen, Bo-Chao Cheng and Cheng-Shong Wu
Department of Electrical Engineering
National Chung-Cheng University
160 San-Hsing, Min-Hsiung, Chia-Yi 621, Taiwan
Email: {pat,huan}@ee.ccu.edu.tw and {bcheng,ieecsw}@ccu.edu.tw
Tel: +886-5-2720411 ext 23253
Fax: +886-5-2720862

***Abstract*— DiffServ services have been supported by many network systems to provide the preferential treatments to users with three service classes: Expedite Forwarding (EF), Assured Forwarding (AF), and Best Effort (BE) services. One prominent high speed network system which has enabled the DiffServ is the Multi-Protocol Label Switching (MPLS) based network. However, current DiffServ mechanism in MPLS network, either using E-LSP or L-LSP, classifies all packets in a LSP as the same class. Such design discipline can not solve the myth of tradeoff between cost and service quality. In this paper, we propose two novel service models, AF^+_{non-RA} and AF^+_{RA}, to enable the E-LSP and L-LSP to tradeoff between the two opposing objectives. The major technical challenge involves how to meet each user's SLA requirement while maximizing system revenue. AF^+_{non-RA} and AF^+_{RA} services are modeled as two-dimension and one-dimension birth-death processes respectively to evaluate their performance under different traffic scenarios. The 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-RA} and AF^+_{RA} meet the cost-effective requirements and provide a practical solution to DiffServ-based MPLS Networks.**

I. Introduction

With the success of Internet and advances of broadband access technologies, the number of multimedia users is increasing rapidly over the past few decades. To take the advantages of the resource on the Internet, users access broadband networks through various technologies such as xDSL, Cable Modem, Fiber-to-the-Home (FTTH), and Digital Wavelength Division Multiplexing (DWDM). One way for these technologies to accommodate users' resource demanding, thus to guarantee their quality of service (QoS), is to employ new infrastructure to increase bandwidth; 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.

VoIP is one of the most popular applications [1] over the broadband access networks due to its low cost compared to conventional voice service via PSTN. Inherited from its interactive and real-time nature, it is very sensitive to network congestions and it requires more stringent QoS than data traffic. For VoIP applications, some QoS requirements shall be maintained to make them attractive to users [2], for example, good voice quality and short call waiting time. The former QoS metric can be achieved by employing higher bit rate codec but it needs more bandwidth; the later QoS metric can be improved by giving higher data forward priority to this application.

Internet and network communities have addressed the QoS issues about multimedia for a long time. The emergence technology known as "the Differentiated Service Model over Multi-protocol Label Switch" (Diff-Serv/MPLS) [3] can enhance QoS provisioning ability for the conventional IP-based networks. The DiffServ-based MPLS network is a suitable platform to run VoIP applications, since it can support differentiated traffic classes and provide preferential treatments to users. In addition, such networks 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 thus lend itself very well to the VoIP applications.

However, the conventional service providers support only either one of the three service classes in a single VoIP trunk to a group of VoIP subscribers. Under such service framework, it is difficult for users to determine which service class to sign up will do themselves the best. To help users to receive appropriate services, we will propose two novel service models in this paper, each with different service level agreement (SLA) signing policies for comparison. The analytical results in this paper will provide useful information to both users and service provider for signing a cost-effective contract which provides a better trade off between cost and forwarding priority.

The remaining of this paper is organized as follows. In section II, Differentiated Service model and DiffServ-based MPLS networks are briefly reviewed. In section III, we propose a new AF^+ service model which is followed by its two variants : 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 section III-A and section III-B, respectively. Performance analysis for both models are illustrated in section IV. Finally, section V concludes this paper.

II. Background

The Differentiated service model supports three services, listed in their priority orders: the *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. 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, core networks does not provide AF users any "hard" QoS guarantees, but "soft" ones. In other words, 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 system is not congested, or pay less to receive lower priority service and thus get much poorer throughput than expected.

In literature, the EXP-Inferred-PSC LSPs (E-LSP) and Label-Only-Inferred-PSC LSPs (LLSP) 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 can not solve the myth of trade-off between cost and service quality. Additionally, the ideal of AF^+ service model was presented in [10] first. To provide cost-effective DiffServ solutions in MPLS networks, we proposed two service models, AF^+_{non-RA} and AF^+_{RA}, to extend the conventional AF service for non-rate adaptive and rate adaptive call admission policies respectively.

III. System and Service Models

The conventional 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. Consider a VoIP user requests an EF service when the VoIP trunk contracted previously is congested. A service provider can still get some profits by providing a degraded service (AF) for a discount price, rather than just request this EF service request. The conventional service model do not support such functionality, which motivates us to propose the following generic AF^+ service model.

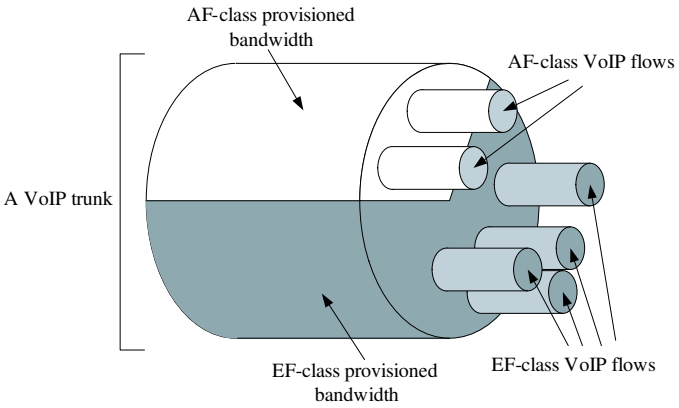

Fig. 1. EF-class and AF-class provisioned bandwidth in a VoIP trunk.

In this section, we first propose a generic AF^+ service model which is constructed based on the DiffServ service but with an adaptive SLA signing policy. 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 Diff-Serv/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. In a single VoIP trunk, as an example illustrated in Fig. 1, the first coming VoIP flows receive EF service up to m flows and the rest of n flows receive only AF class service. 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 a VoIP trunk congestion threshold and can be expressed as

$$\alpha = \frac{m \cdot BW_{EF}}{BW_{total}};\ 0 \leq \alpha \leq 1. \tag{2}$$

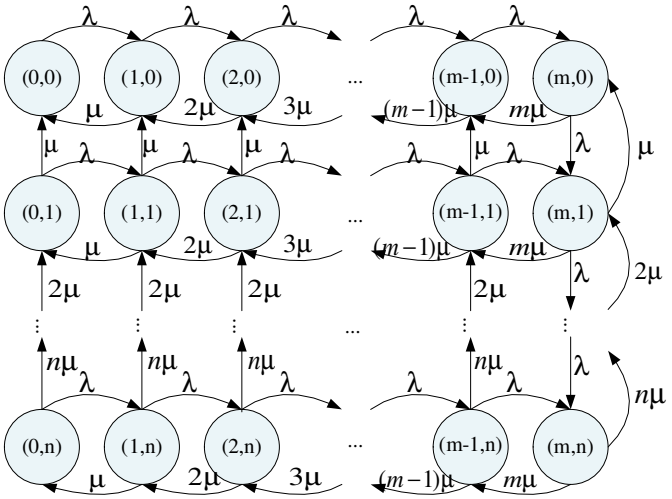

Fig. 2. Two-dimension state-transition-rate diagram for the AF^{+}_{non-RA} service model.

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 IV. In the following paragraphs, we introduce and analyze its two variants, denoted as AF^{+}_{non-RA} and AF^{+}_{RA}, respectively. They differs in their SLA policies. AF^{+}_{non-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.

A. AF^{+}_{non-RA} Service Model

AF^{+}_{non-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 request will be treated as AF class up to n calls. But in this AF^{+}_{non-RA} model, the AF class service, once determined by Diff-Serv 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, management the states of VoIP sessions and decide which service class they are before VoIP connection is established.

The proposed AF^{+}_{non-RA} service can be modeled as a two-dimension birth-death process [11]. Fig. 2 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 $Erlang = \lambda/\mu$.

Observe that in Fig. 2, 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 $0 \le i \le m$ and $0 \le j \le n$, respectively. Let $P_{i,j}$ be the stationary sate probability of (i, j), it can be found by solving the equilibrium equations as below.

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

The normalization equation for the above equilibrium equations is

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

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_{blk} can be derived via Eq. (5).

$$
P_{blk} \quad = \quad P_{m,n} \tag{5}
$$

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

$$
U_{EF} \quad = \quad \frac{1}{m} \sum_{j=0}^{n} \sum_{i=0}^{m} i P_{i,j} \tag{6}
$$

$$
U_{AF} \quad = \quad \frac{1}{n} \sum_{j=0}^{n} \sum_{i=0}^{m} j P_{i,j} \tag{7}
$$

The cost function of AF^{+}_{non-RA} service model is defined as below.

$$
\begin{aligned}
J \quad &= \quad \beta(1 - U_{EF}) + (1 - \beta)(1 - U_{AF}) \\
&= \quad 1 - \beta U_{EF} - (1 - \beta) U_{AF}
\end{aligned} \tag{8}
$$

where β is a penalty ratio. Penalties are incurred by bandwidth waste. Due to the different cost of EF-class and AF-class resources, the proportion of both penalties is $\beta : (1 - \beta)$.

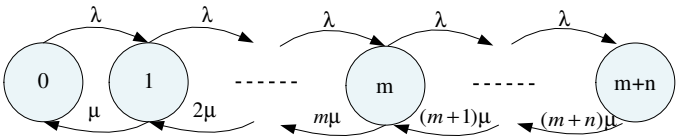

Fig. 3. Two-dimension state-transition-rate diagram for the AF^+_{RA} service model.

B. AF^+_{RA} 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 management the states of VoIP sessions anytime. If there is available EF-class bandwidth, the a 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 [11]. Fig. 3 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.

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

where

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

The normalization equation for the above equilibrium equations is

$$\sum_{k=0}^{m+n} P_k = 1 \quad (10)$$

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} can be derived via Eq. (11) that is alike to Erlang-B Model.

$$P_{blk} = P_{m+n} \quad (11)$$

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

$$U_{EF} = \frac{1}{m} \left(\sum_{k=0}^{m} k P_k + \sum_{k=m+1}^{m+n} m P_k \right) \quad (12)$$

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

The cost function of AF^+_{RA} service model is alike to which of AF^+_{non-RA} service model shown in Eq. 8.

IV. Numerical Result

The AF^+_{non-RA} model is a two-dimensional Markov chain with a complex structure, so performance metrics of a closed-form solution cannot be derived. The sophisticated commercial optimization tool, LINGO, is then applied to solve it. However, the AF^+_{RA} model degenerates to a one-dimensional Markov chain and is easily analyzed. This section presents numerical results for both service models. The performance metrics of each model include: the call blocking rate and the resource utilization.

A. Call Blocking Rate

We assume that the traffic intensity of VoIP calls equals to 25 ($Erlang = 25$). The call blocking rates of AF^+_{non-RA} and AF^+_{RA} service models are calculated from Eq. 5 and Eq. 11, and the results are illustrated in Fig. 4. The Y axis is the call blocking rate, the X axis is the total provisioned bandwidth $m+n$, and the Z axis is the ratio α of the total provisioned bandwidth.

In the figure, we can observe one important phenomenon. First, the call blocking rates are depend on the total provisioned bandwidth $m+n$ regardless of the congestion threshold α. If the silence suppression of VoIP traffic is ignored, both of AF^+_{non-RA} and AF^+_{RA} service models are alike to Erlang-B model. Thus, the total bandwidth can be assigned to tune the alternative issue of call blocking rate and provisioned bandwidth cost that can refer to Erlang-B model in the conventional telecommunication services.

B. Optimal Solution

In this analysis, we assume that the total bandwidth of 40 channels ($m+n = 40$) is reserved for AF^+_{non-RA} and AF^+_{RA} service models, and both of their call blocking rates are guaranteed to less than 0.0012 while the Erlang is less than 25. The EF-class bandwidth utilizations of AF^+_{non-RA} and AF^+_{RA} service models are calculated from Eq. 6 and Eq. 12, and their AF-class bandwidth utilizations both are calculated from Eq. 7 and Eq. 13.

The relationships of bandwidth utilization and bandwidth provisioning ratio are illustrated in Fig. 5. We observe two phenomenons. First, if a small α (such as $\alpha \to 0$) is selected, both EF-class bandwidth utilizations of AF^+_{non-RA} and AF^+_{RA} are greater than 0.96. Otherwise, if a large α (such as $\alpha \to 1$) is selected, both of them are decreasing, because the EF-class bandwidth is enough to serve the most VoIP calls. Second, the EF-class bandwidth utilization of AF^+_{RA} is lager than that

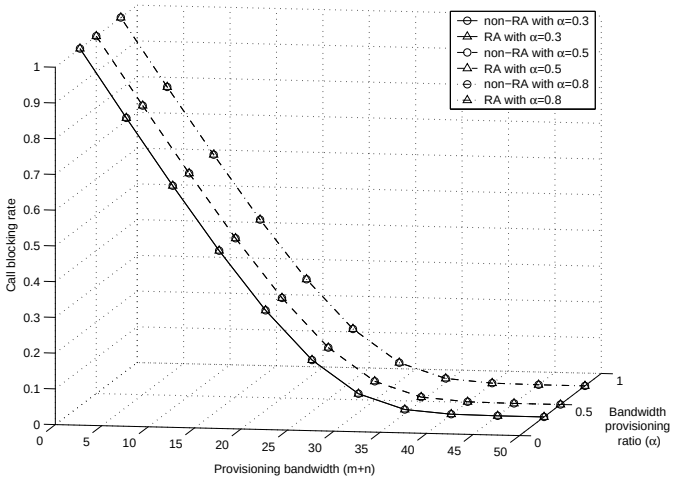

Fig. 4. The call blocking rates of AF^{+}_{non-RA} and AF^{+}_{RA} service models while $Erlang = 25$.

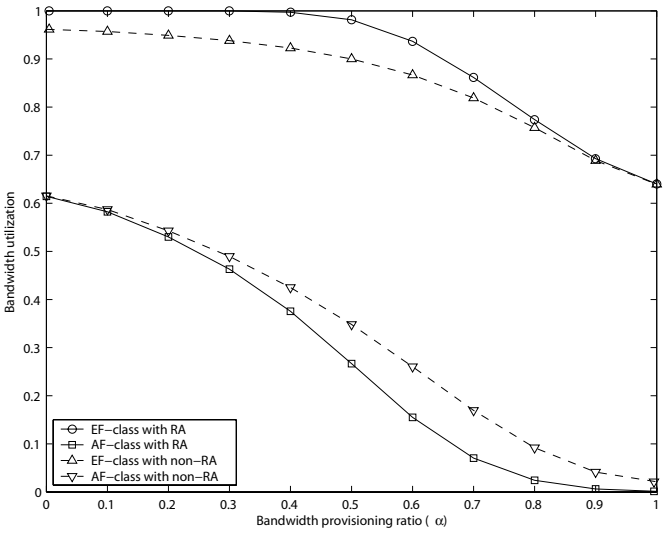

Fig. 5. The comparison of provisioned bandwidth utilizations with AF^{+}_{non-RA} and AF^{+}_{RA} service models while $m + n = 40$.

of AF^{+}_{non-RA}, because of the AF service could be upgraded to EF service in AF^{+}_{RA} service model while the EF-class bandwidth is available.

The optimal solutions are illustrated in Fig. 6. If we only consider the cost issue, $\alpha = 0$ is the best solution for both service models. Only AF-class bandwidth is provisioned. However, we have alternative issue of QoS and cost in fact. If the penalty ratio β is well-known, we could select the maximal α and that has the minimal cost. For example, according to the consideration of QoS and cost, we should select $\alpha = 0.3$ for AF^{+}_{RA} service model with $\beta = 0.99$ and $\beta = 0.999$. Besides, in Fig. 6, we can observe that AF^{+}_{RA} is the better service policy regardless of the complexity of implementation.

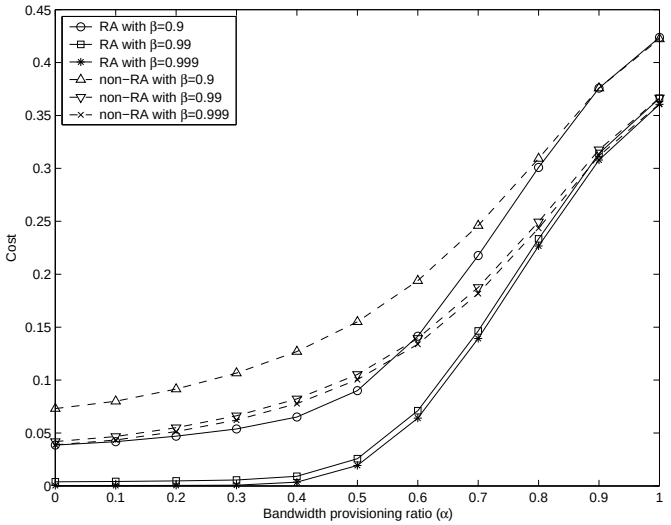

Fig. 6. The comparison of costs with AF^{+}_{non-RA} and AF^{+}_{RA} service models while $m + n = 40$.

V. Conclusion

Tradeoff between 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-RA} and AF^{+}_{RA} service models to enable E-LSP and L-LSP to accommodate emerging applications such as VoIP and real time video streaming over the Diff-Serv/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-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. We modeled AF^{+}_{non-RA} and AF^{+}_{RA} services as two-dimension and one-dimension birth-death processes respectively to evaluate their performance under different traffic scenarios.

The 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-RA} and AF^{+}_{RA} meet the cost-effective requirements and provide a practical solution to DiffServ-based MPLS Networks.

Acknowledgement

We would like to thank the anonymous reviewers for their valuable efforts and comments. This research was supported in part by the National Science Council (NSC) in Taiwan under the grant number NSC-94-2219-E-194-002 and NSC-94-2219-E-194-004.

REFERENCES

[1] B. Goode, "Voice over Internet protocol (VoIP)," *Proceedings of the IEEE*, vol. 90, pp. 1495–1517, Sept. 2002.

[2] V. Fineberg, "A practical architecture for implementing end-to-end QoS in an IP network," *IEEE Communications Magazine*, vol. 40, pp. 122–130, Jan. 2002.

[3] 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.

[4] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An Architecture for Differentiated Services," RFC 2475, Dec. 1998.

[5] E. Rosen, A. Viswanathan, and R. Callon, "Multiprotocol Label Switching Architecture," RFC 3031, Jan. 2001.

[6] K. Muthukrishnan and A. Malis, "A Core MPLS IP VPN Architecture," RFC 2917, Sept. 2000.

[7] F. L. Faucheur and W. Lai, "Requirements for Support of Differentiated Services-aware MPLS Traffic Engineering," RFC 3564, July 2003.

[8] 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, Mar. 2002.

[9] J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski, "Assured Forwarding PHB Group," RFC 2597, June 1999.

[10] 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," in *Proc. IEEE VTC 2004-Fall*, Sept. 2004.

[11] L. Kleinrock, *Queueing Systems; Volime 1: Theory*. New York, USA.: John Wiley & Sons, 1972.