Preemption Policy in QoS-Enabled Networks: A Customer Centric Approach

Iftekhar Ahmad and Joarder Kamruzzaman

Gippsland School of Computing and Information Technology
Monash University, Victoria 3842, Australia
Email Address: Iftekhar.Ahmad@infotech.monash.edu.au
Phone: +61 3 5122 6135
Fax: +61 3 9902 6842

Preemption is a key technique to provide available and reliable services to high priority connections in a QoS-enabled network. Preemption technique is governed by a policy which makes the decision about which connections to preempt when resource scarcity is experienced. Existing preemption policies consider preemption rate, priority of connection and preempted bandwidth as the deciding criteria. In this paper, two additional criteria, namely, user satisfaction and revenue index are introduced in formulating an objective function that defines preemption policy as an optimization problem. We propose a model for the network enterprise to calculate an estimated level of user satisfaction of preemption enabled call connections in the context of service continuity and incorporate this information into preemption policy. Simulation results show that the proposed policies achieve improved customer perceived satisfaction and revenue index compared to the existing policies while preserving good system utilization, preemption rate and priority level of preempted calls. The improved user satisfaction and revenue index indicate higher prospects of revenue return and make the proposed policies highly attractive to the network providers as well as beneficial to the customer.

Key Words: Preemption, quality of service (QoS), service continuity, customer satisfaction, revenue index.

ACM Classification: C.2.3 Network Operations

1. INTRODUCTION

Bandwidth reservation and management have been a key issue for ensuring quality of service (QoS) in both wireless and wireline networks for years due to increasing demand of multimedia and distributed applications. Multimedia applications over the Internet like video conferencing, video on demand, live broadcast of TV programs, medical applications like remote surgery or telemedicine, tele-teaching and distance learning etc. require point to point guarantees of QoS. Assurance of QoS is equally crucial for applications like grid computing, distributed simulations that run on multiple super computers, tele-immersion applications that require simultaneous access to databases, CAD tools and rendering devices. Differentiation of call connection is often very important in a QoS-Enabled network to ensure connection specific quality of service. One of the

Manuscript received: 5 August 2005
Communicating Editor: Sidney A. Morris

widely used techniques to ensure connection specific QoS is the ‘preemption of call connection’ to supply enough resources to high priority call connections so that their QoS is rightly maintained. Preemption of connections is required to make resources available in response to resource scarcity for connections having relatively high importance which is often indicated by proper priority level. Preemption is also important to maintain QoS if over-provisioning of resources is experienced due to any non-stationary network condition like failure of links or nodes and a priori unknown traffic pattern.

The preemption technique has been studied and used in a number of previous research works. Yao, Mark, Chew, Lye and Chua (2004) found higher gain in system utilization when preemption rules were applied in virtual partitioning (VP) resource allocation for multi-class traffic in cellular systems with QoS constraints. VP behaves like unrestricted sharing of resources under light traffic load. When the load becomes heavy, partitioning of resources is strictly applied and this is achieved by preemption of overloaded traffic classes which have been using resources beyond their nominal partitioning limit. Iera, Molinaro and Marano (2000) proposed a traffic management strategy for integrated ATM-satellite systems. This work used the statistical multiplexing of the traffic sources and mainly exploited dynamic resource management based on a preemption policy. Their results showed that the preemption policy when used for dynamic resource management provided better satellite bandwidth usage while provisioning QoS to both real-time and non real-time VBR traffic. Do, Park and Lee (2002) proposed a dynamic priority adjustment (DPA) control as a channel assignment policy for a multiclass multicode CDMA system with guaranteed QoS. The DPA system employed the preemption technique to ensure QoS for real time traffic classes over non real time traffic classes with a restricted preemptive priority. High channel utilization and guaranteed QoS were observed in the DPA control system used for channel assignment. Ahmad, Kamruzzaman and Aswathanarayaniah (2004) used preemption rules to make resources available for Book-Ahead (BA) calls which reserve the resources in advance for their time sensitive heavy traffic load. Resources from Instantaneous Request (IR) calls are preempted to ensure the QoS of BA calls which are mainly time sensitive real time multimedia or distributed applications. The preemption technique is equally attractive for ATM, MPLS and IP based QoS-enabled networks (Oliveira, Scoglio, Akyildiz and Uhl, 2002).

The preemption technique is governed by a preemption policy that determines which call connections to preempt under resource scarcity. For its high importance, an optimal preemption policy has attracted increasing interest from researchers over a period of time. Garay and Gopal (1992) addressed the call preemption selection problem in a centralized network environment. In this work it was shown that the process of selecting which calls to preempt in a centralized environment with an objective to minimize the number of calls to be preempted or to minimize the amount of bandwidth to be preempted is a NP complete problem. They presented a heuristic to avoid the computational intractability. However, most of the resource reservation protocols proposed in the recent past like RSVP, RSVP-TE or ATM signaling use decentralized computation where each node has to make decisions and performs functions independent of the other control points. Considering a decentralized architecture, Peyravian and Kshemkalyani (1997) proposed two algorithms: Min_Conn and Min_BW, which are computationally tractable to find the calls to be preempted in a decentralized architecture. The Min_Conn algorithm first minimizes the number of call connections, then searches the combination of connections to minimize the bandwidth and if there is a choice of such combinations, it selects a combination that has the least priority for the call connections to be preempted. The Min_BW algorithm finds a solution in the order of importance of bandwidth, priority and number of connections. Oliveira *et al* (2002) further improved the

Min_Conn and Min_BW algorithms by formulating an objective function to minimize whose parameters can be adjusted by the service provider in order to give importance to the desired criteria. The criteria they considered for a preemption policy were number, priority and bandwidth of preempted calls. However, service continuity of calls which is perceived by users as of high importance in a QoS-enabled network (Campanella, Chivalier and Simar, 2001) has not been considered in any of the previous works. When preemption becomes inevitable, service continuity of preempted calls is disrupted which leads to user dissatisfaction. The objective of this paper is to introduce new optimization criteria based on user satisfaction in the context of service continuity along with the three previously proposed optimization criteria in the literature. A method to estimate user satisfaction for preemption policy is proposed. We introduce Revenue Index (RI), a parameter used in strategic revenue analysis taking user satisfaction into consideration, to relate to the criteria of the objective function and show that it serves as an important metric to find a balanced importance of all the criteria. We also present a formulation of the objective function incorporating the RI index which yields higher user satisfaction and revenue index. This work concentrates on user satisfaction in the context of service continuity with the assumption that classical bandwidth reservation schemes with service level agreement (SLA) take care of other potential sources (e.g., packet loss, latency, service cost etc.) of user dissatisfaction.

The rest of the paper is organized as follows: in Section 2, we introduce the problem formulation. Existing preemption policies to address the problem is discussed in Section 3. In Section 4 we propose a model to calculate user satisfaction. Section 5 includes the proposed policies. Section 6 shows the simulation result followed by concluding words at Section 7.

2. PROBLEM FORMULATION

Resource reservation is a widely recommended technique which ensures point to point QoS guarantees to the end applications in communication networks. MPLS uses RSVP-TE for integrated services and DS-TE for differentiated services to reserve bandwidth and thereby maintain QoS in MPLS network (Awduche, 1999; Faucheur, 2001). ATM signaling (PNNI draft, 1995) and RSVP (Braden, Zhang, Berson, Heroz, and Jamin, 1997) are the other protocols used for bandwidth reservation and QoS management in ATM and TCP/IP based networks respectively. RSVP-TE and DS-TE support preemption of lower priority LSPs with a preemption enable attribute in the case of resource scarcity experienced due to the arrival of a LSP with higher priority in a MPLS network. An ATM network also supports preemption of low priority calls to make room for high priority calls. Although preemption is not a mandatory attribute in an IP-based network, it has become a very important scheme in both Integrated and Differentiated Service architectures in recent times (Oliveira *et al*, 2002). In integrated architecture, per call resource reservation is required and preemption here is particularly important in context of Book-Ahead (BA) reservation (Ahmad *et al*, 2004; Greenberg, Srikant and Whitt, 1999). A BA call is required to declare its bandwidth demand and duration well before its actual starting time. Provision of resources is ensured at its starting time because the BA application is time sensitive and highly resource consuming. At the starting time of BA applications if resource scarcity is evident, the preemption technique is applied to drop some of the on-going Instantaneous Request (IR) calls (Figure 1). IR reservation is made immediately after the call acceptance while in BA reservation availability of resources is confirmed well ahead of usage time. This is why BA calls always enjoy higher priority at resource contention. In differentiated architecture traffic is considered per class basis and preemption is required to make room for QoS sensitive traffic class in case of resource scarcity.

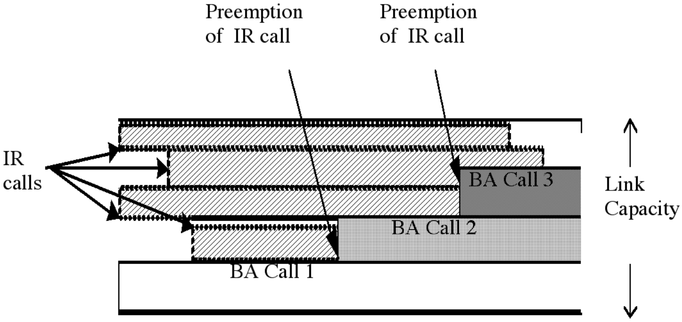


Figure 1: Preemption in BA reservation

2.1 Problem Statement

Consider a connection request i, with bandwidth requirement b_i and priority y_i. If $\sum b_{j\in S} + b_i > C$, then find a set $U \subseteq S$ such that $y_{j\in U} < y_i$, $\sum b_{j\in U} \geq \sum b_{j\in S} + b_i - C$ where S is the set of call connections currently using the link and C is the link capacity. All the elements of set U have the attribute 'preemption enabled'.

3. EXISTING PREEMPTION POLICIES

A preemption policy finds a set U in response to resource scarcity. A solution which closely fits with the problem statement is proposed by Oliviera *et al* (2002). In their work a mathematical formulation was proposed which combined the interest of three important objectives: i) minimize the priority of preempted calls, ii) minimize the number of preempted calls, and iii) minimize the preempted bandwidth. The first objective is important as it is trivially objectionable if calls with higher priorities are disadvantaged over calls with lower priorities in a preemption policy. If preemption is inevitable, lower priority calls should be preempted first. The second objective is equally important as higher numbers of preemptions result in more disruption of service continuity. After preemption the preempted call may be tried for possible re-routing. However, high speed networks often apply a "no re-routing" restriction because of the higher bandwidth-delay product which far exceeds the practical limit of buffer size (Awerbuch, Azar and Plotkin, 1993). Re-routing is also highly expensive in a large network as resource reservation of the preempted call is needed to be confirmed over the new route which can be quite long. Moreover, there is no guarantee that another route that satisfies the QoS of the preempted call will be available immediately after preemption. Disruption of service continuity of the preempted call is thus highly probable in such cases. Lowering the preemption rate is a good solution to avoid the disruption of service continuity. The third objective is to minimize the preempted bandwidth. It guarantees minimal wastage of resources and improves system utilization. The three key interests were then combined into a single objective function which was a weighted sum of the above mentioned criteria. Mathematical formulation of this policy is given as (Oliveira *et al*, 2002)

$$\mathrm{F}(\mathbf{z}) = \alpha\,(\mathbf{z}\,.\,\mathbf{y}^{\mathbf{T}}) + \beta\,(\mathbf{z}.\,\mathbf{1}^{\mathbf{T}}) + \gamma\,(\mathbf{z}\,.\,\mathbf{b}^{\mathbf{T}}) \tag{1}$$

The vector **z** is an optimization variable and is composed of n binary variables where n is the total number of on-going call connections in the system. Each element $\mathbf{z}(l)$ of **z** is defined as

$$\mathbf{z}(l) = \begin{cases} 1 & \text{if call } l \text{ is preempted} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Term $\mathbf{z.y}^T$ represents the priority of the preempted calls, $\mathbf{z.1}^T$ represents the number of preempted calls and $\mathbf{z.b}^T$ represents the total preempted bandwidth. Vectors **y** and **b** indicate the priority and bandwidth of the existing call connections respectively where an increasing value of y_i indicates higher priority. Terms α, β and γ are the weights that select the level of preference. The solution of the problem stands as to minimize the objective function $F(\mathbf{z})$ subject to the following constraint

$$\mathbf{z.b}^T > r \text{ where } r = \sum b_{j\in s} + b_i - C \tag{3}$$

Online use of this kind of optimization is feasible in small and medium size networks. It provides fast and accurate results. However, for a large size network online use of optimization may prove infeasible in consideration of computation and time complexity. A heuristic which follows the optimization results is then a good choice for a large size network. Oliviera *et al* (2002) proposed a heuristic for a large size network as

$$H(l) = \alpha\, y(l) + \beta + \gamma\,(b(l) - r)^2 \tag{4}$$

where y(l) indicates the loss in priority and b(l) indicates the bandwidth of the call connection l. The term $(b(l)–r)^2$ is used to ensure that less number of connections are placed for preemption. Function H is calculated for each call and the calls are preempted in order of increasing value of H. In their work Oliviera *et al* (2002) assumed that bandwidth was available in bandwidth modules and on that basis an integer optimization approach was followed to solve the problem.

In the context of BA reservation, Greenberg *et al* (1999) proposed to preempt the IR calls in Last In First Out (LIFO) fashion. They argued that if the call with the most recent arrival time was preempted the impact on the successfully transmitted amount of data would be minimal.

4. MODELING CUSTOMER SATISFACTION

One of the novelties of this paper is to model and introduce a new criterion, user dissatisfaction in the context of service continuity into the objective function. The motive is to minimize the loss of user satisfaction which occurs due to disruption of service continuity. According to the current study (Campanella *et al*, 2001; Hardy, 2001), service continuity is considered as a very important issue for a QoS aware application. Most of the users prefer uninterrupted service and perceive service continuity as one of the important QoS guarantee metrics (Campanella *et al*, 2001). A disruption of service continuity is highly probable for a preempted call with the existing preemption policies.

Importance of service continuity on user satisfaction is different for different applications. For an application which performs an atomic task (e.g., bank transaction) over its complete duration, the utility gain is zero unless it is fully complete. An application whose importance increases sharply towards the end of its completion (e.g., live broadcasting of a game or a movie in video on demand) provides more satisfaction towards the end of its duration and thus the satisfaction curve is exponential in nature. Applications like guaranteed data transfer or voice conversation have different satisfaction curves. We define the user satisfaction (US) function (Figure 2) as:

$$US_i = \begin{cases} e^{k\left(\frac{T_i}{D_i}-1\right)} & \text{if } T_i < D_i \\ 1 & \text{if } T_i \geq D_i \end{cases} \tag{5}$$

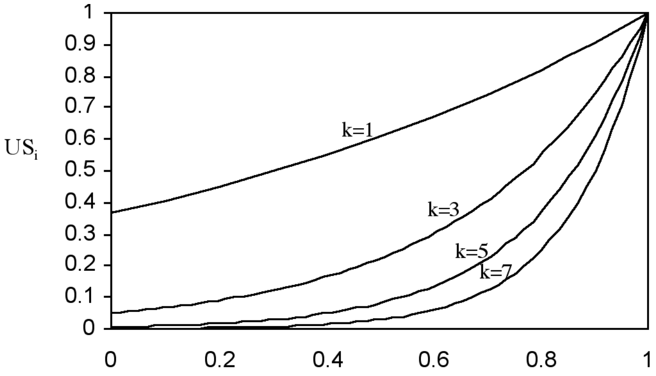


Figure 2: Lever of user satisfaction for different values of k

where T_i is the time of data transmission before preemption and D_i is the complete data transmission time of call connection *i* if not placed for preemption. However, when a call connection enters into the system, the network provider does not have an idea of the exact value of D_i until it finishes and thus the network provider can at best estimate the lifetime when the connection is active. In this work, for an active connection we model D_i as equal to the mean data transmission time calculated from the observed lifetime of the calls of similar type (group) to which connection *i* resembles the most (e.g., voice, video, or ftp type of application). The value of *k* is application specific and it indicates the emphasis of service continuity on user satisfaction. For example, the value of *k* should be higher for applications like telecast of sports over the Internet for which the final result carries the highest significance. Term D_i can be different for different groups of applications and it can be obtained from distributions of applications in real time networks (Lin, Chang and HSU, 2002). US_i denotes the estimated level of user satisfaction when calculated on the network provider's side (calculated based on the estimated value of D_i and used in the preemption policy) whereas it shows the actual level of user perceived satisfaction when calculated on the users' side (calculated based on the exact value of D_i for call i as reported in Section 6). Since user satisfaction is a key element in QoS networks, in the following Sections we present new formulations of the preemption policy to minimize the estimated level of user dissatisfaction in addition to the other three criteria.

5. PROPOSED PREEMPTION POLICIES

We propose three different preemption policies which are essentially based on the user satisfaction criterion in addition to the other three criteria and then make comparative studies of those along with the existing preemption policies.

5.1 Preemption Policy without Price Category

In this policy we consider user satisfaction irrespective of the revenue that an individual customer provides. The new objective function is given as (Ahmad *et al*, 2005)

$$F(\mathbf{z}) = \alpha(\mathbf{z}.\mathbf{y}^{\mathrm{T}}) + \beta(\mathbf{z}.\mathbf{1}^{\mathrm{T}}) + \gamma(\mathbf{z}.\mathbf{b}^{\mathrm{T}}) + \delta(\mathbf{z}.\mathbf{d}^{\mathrm{T}}) \qquad (6)$$

Here, **d** is a vector composed of *n* elements for *n* on-going calls and represents the level of user dissatisfaction if the call is preempted. Each element of d_i of **d** is given as

$$d_i = 1 - US_i \tag{7}$$

The complete problem in mathematical formulation is given as

Given $\alpha, \beta, \gamma, \delta, \mathbf{y}, \mathbf{b}, \mathbf{d}$, find $\mathbf{z}$ that minimizes $F(\mathbf{z})$, subject to $\mathbf{z}.b^T > r$.

This is a mixed integer optimization problem. Computational complexity may restrict the use of such on-line optimization in a large network. We propose a heuristic to follow the trend of this optimization. The heuristic used in our study is given by the following equation

$$H(l) = \alpha\, y(l) + \beta + \gamma\,(b(l) - r)^2 + \delta\,(b(l) - r)^2\, d(l) \tag{8}$$

where $d(l) = 1\text{-}US_l$.

The heuristic algorithm is given as following:

```
Algorithm Preemption ($\alpha,\beta,\gamma,\delta,y,b,d$)
  {
     preempt = 0
     Calculate $r = \sum b_{j\in s} + b_i - C$
     for each element $l$ in set $S$ calculate
          $H(l) = \alpha\, y(l) + \beta + \gamma\,(b(l) - r)^2 + \delta\,(b(l) - r)^2\, d(l)$
     while preempt < r
          Select call in increasing order of H
          prempt = preempt + b($l$)
          $\mathbf{z}(l) = 1$
     return z;
}
```

5.2 Preemption Policy with Price Category

Pricing is an important issue for both customers and network providers. A valued customer is likely to enjoy better service and pay more. When the pricing issue is taken into consideration customer satisfaction of valued customers becomes more important. It is thus logical to give attention to satisfaction of customers who pay more. Taking the pricing issue into consideration we propose a new objective function where price per unit bandwidth depends on application type and priority. An application with higher priority and higher demand of QoS pays more and dissatisfaction of the customer using this type of application will cost the network providers a higher loss in revenue earning. With pricing information, the new objective function is stated as

$$F(\mathbf{z}) = \alpha(\mathbf{z}\,.\,\mathbf{y}^{\mathbf{T}}) + \beta(\mathbf{z}\,.\,\mathbf{1}^{\mathbf{T}}) + \gamma(\mathbf{z}\,.\,\mathbf{b}^{\mathbf{T}}) + \delta(\mathbf{z}\,.\,\mathbf{w}^{\mathbf{T}}) \tag{9}$$

where $\mathbf{w}$ is a vector composed of n elements and each element indicates the weighted dissatisfaction. Normalized pricing is the weighting factor for each element. Individual element w_i is calculated as

$$w_i = \frac{p(i)}{\sum_{j=1,\cdots,m} p_j}(1 - US_i) \tag{10}$$

where P_j is the price per unit bandwidth of the j-th price category, *m* is the total number of different price categories and p(i) is the price per unit bandwidth of the price category to which call connection *i* belongs.

The heuristic for the mixed optimization problem is given as

$$H(l) = \alpha y(l) + \beta + \gamma\, (b(l) - r)^2 + \delta\, (b(l) - r)^2\, w(l) \tag{11}$$

where $$w(l) = \frac{p(l)}{\sum_{j=1,\cdots,m} p_j}(1 - US_l) \tag{12}$$

5.3 Preemption Policy with Revenue Index

Revenue return is one of the main driving forces for a network provider. In economics, the prospects of revenue earning are often determined by the level of user satisfaction. User satisfaction is thus highly crucial specially when the long term future of the enterprise is considered. In a recent study, Lewis (2002) proposed a measurement of the relationship between customer satisfaction and revenue prospects. A metric called the Revenue Index (RI) was proposed by Lewis that reflects the relationship between customer satisfaction and revenue return. A two step calculation of the RI index was proposed as follows (Lewis, 2002):

Step 1: Calculate the percentages of each of the four satisfaction groups of survey respondents: (i) Totally satisfied (ii) Somewhat satisfied (iii) Somewhat dissatisfied and (iv) Totally dissatisfied.

Step 2: Multiply those percentages of the four categories by the weighting factors. The weighting factors are obtained using the multivariate linear regression over surveyed data. RI is calculated as follows:

RI = 1.0 × % of totally satisfied respondents + 0.38 × % of somewhat satisfied respondents + 0.068 × % of somewhat dissatisfied respondents – 1.80 × % of totally dissatisfied respondents

The rationale of such a calculation is based on the observation (Lewis, 2002) that a fully satisfied customer pays 100% of revenue for the specific product or service. A somewhat satisfied customer pays 38% of the revenue that a fully satisfied customer pays. A somewhat dissatisfied customer pays 6.8% while a fully dissatisfied customer subtracts 180% of the revenue. The numerical figures were found from the relationship that emerged between customer satisfaction and revenue earning based on practical data collected over a long period of time.

A healthy RI is certainly a strong goal for a network enterprise, but no work known to the authors has yet focused on modeling and targeting RI in communication networks. In this work we used the formulation proposed by Lewis to map estimated user satisfaction to Revenue Index. The calculated RI is then used as one of the criteria for preemption policy. We propose to minimize the weighted loss in RI in addition the other three criteria and the objective function is formulated as

$$F(\mathbf{z}) = \alpha(\mathbf{z}.\mathbf{y}^{\mathbf{T}}) + \beta(\mathbf{z}.\,\mathbf{1}^{\mathbf{T}}) + \gamma(\mathbf{z}.\mathbf{b}^{\mathbf{T}}) + \delta(\mathbf{z}.\mathbf{x}^{\mathbf{T}}) \tag{13}$$

where **x** is a vector composed of *n* elements and each element indicates the estimated weighted loss in RI per call basis. Each element x_i of **x** is calculated as

$$x_i = \frac{p(i)}{\sum_{j=1,\ldots,m} p_j}(1-R_i) \tag{14}$$

where R_i indicates the level of revenue index for call connection *i* if it is placed for preemption. The heuristic for the above optimization problem is stated as

$$H(l) = \alpha\, y(l) + \beta + \gamma\,(b(l) - r)^2 + \delta\,(b(l) - r)^2\, x(l) \tag{15}$$

where $x_i = \frac{p(i)}{\sum_{j=1,\ldots,m} p_j}(1-R_i)$ (16)

For different price category applications, the overall RI shown in Section 6 is calculated in the integrated form as

$$RI = \frac{p_1}{\sum_{i=1,\cdots,m} p_i} RI_1 + \frac{p_2}{\sum_{i=1,\cdots,m} p_i} RI_2 + \cdots\cdots + \frac{p_m}{\sum_{i=1,\cdots,m} p_i} RI \tag{17}$$

where RI_i indicates the revenue index for call connections belonging to the *i*-th price category measured at the users' end.

6. SIMULATION RESULTS

Simulation of the proposed policies has been done in the context of Instantaneous Request (IR) and Book-Ahead (BA) reservation for multimedia traffic. Book-ahead reservation requires a guarantee of resource availability in advance and enjoys higher priority over an IR call. A preemption policy plays a very important role when a BA connection becomes active and requires resources to be preempted in a scenario (Figure 1) where resources are shared between IR and BA call connections (Ahmad *et al*, 2004; Greenberg *et al*, 1999). A single bottleneck topology used for the simulation remains the same as in a number of related research works (Ahmad *et al*, 2004; Wang and Schulzrinne, 2000; Yi and Kim, 2002; Lin *et al*, 2002). The capacity of each link is considered as 10 Mbps. IR arrivals to the core link are assumed to follow a Poisson distribution with a mean arrival rate of 11 calls per minute. The arrival of BA calls also follows a Poisson distribution with a mean arrival interval of 50 sec. The bandwidth demand for IR calls is assumed to be exponentially distributed with a mean of 256 kbps. The bandwidth requirement of BA calls is also exponentially distributed with a mean of 1.25 Mbps. To nullify the impact of the difference in mean call holding times, the call duration for both BA and IR calls are determined by exponential distribution with the same mean of 300s. Results in this section are shown for BA limit = 0.8 which physically limits the maximum usage for aggregate BA calls up to 80% of the link capacity. The value of *k* is set depending on the application type and is chosen within the range of (1~8) to represent one of eight applications. In our simulation three priority levels of IR calls are considered, 1 being the highest priority level and 3 being the lowest. BA calls are considered as non-preemptable and are the highest priority calls. Traffic analyses shown in this section are for the core link. Since a multiple bottleneck topology is basically a collection of multiple core links, traffic analysis of a single core link works as the basis. We used a modified version of the ANCLES simulator (ANCLES, 2005) to conduct the simulation.

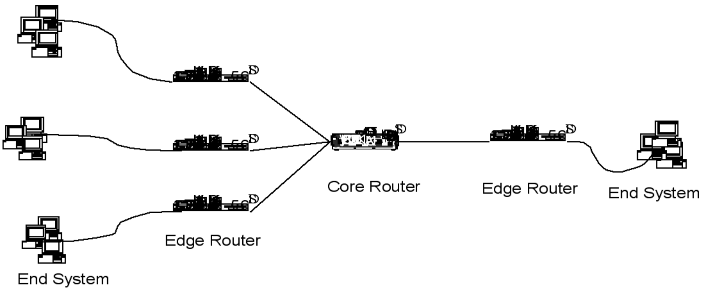


Figure 3: Simulation topology

Results presented in this section investigate a number of important network parameters e.g., user satisfaction, preemption rate, priority level of preempted calls and utilization. The proposed preemption policy is compared against existing preemption policies. It may be noted that in implementing the proposed preemption policy user satisfaction was modeled after the estimated lifetime (D_i) of calls (Eq. 5) whereas the user satisfaction reported in this section is calculated using the actual lifetime (i.e., the lifetime the call would have continued if not preempted).

6.1 Preemption Policy without Price Category

Figure 4 shows the average level of user satisfaction of preempted calls in three different preemption policies: i) Last In First Out, ii) optimization with priority, number and bandwidth criteria (PNB-optimization using Eq. 1), and iii) the proposed optimization with priority, number, bandwidth and estimated user satisfaction (PNBS-optimization using Eq. 6). The mixed integer optimization technique was applied for PNB and PNBS optimization using the standard LINDO API 2.0 tools (Lindo API, 2005). Co-efficient δ associated with criterion user dissatisfaction was varied keeping α, β, γ fixed (=1.0) in order to investigate its effect on performance. The simulation results show that PNBS-optimization achieves the highest average user satisfaction for the preempted calls. Significant improvement in user satisfaction is achieved by PNBS-optimization for all the values of δ. The improvement is higher than 7% over the LIFO policy and 3% over PNB-optimization policy for all

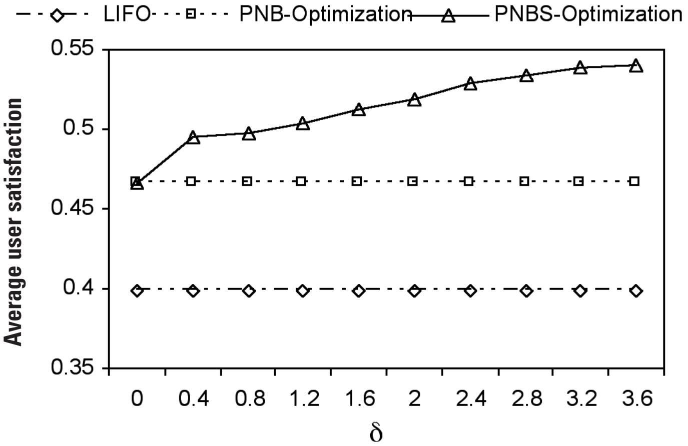


Figure 4: Average level of user satisfaction in different preemption policies

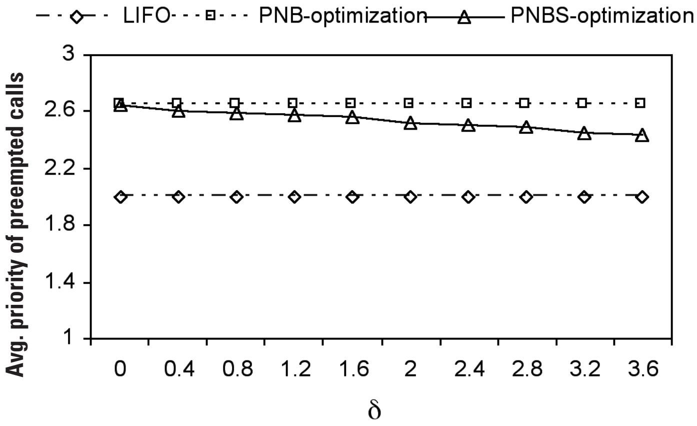


Figure 5: Average priority of preempted calls in different preemption policies

δ >0. As the value of δ is increased the improvement becomes more evident. At δ = 3.6, PNBS-optimization achieves around 14% higher satisfaction over the LIFO policy and 8% higher satisfaction over the PNB-optimization policy. Simulation result for δ>3.6 is not shown in this paper because the network performance maintains the same trend for higher δ values.

When a preemption policy (e.g., PNBS, PNB) is structured as a multi-objective optimization problem, the objective function is optimized as a whole. This results in changing the contribution of the existing component criteria (i.e., priority, number of preemption and bandwidth) when a new criterion (i.e., user dissatisfaction) is added. Increasing a particular co-efficient makes the associated criterion contribute more than the other three criteria to the objective function. This is evident in Figure 5 where PNBS-optimization is found to preempt calls with relative higher priority compared to PNB-optimization for a higher value of δ. The LIFO policy suffers the most as it does not consider priority of calls for preemption. Figure 6 shows the preemption rate in different preemption policies. Low preemption rate is highly desirable in a QoS-enabled network. The PNBS-optimization policy achieves a lower preemption rate than the other two policies. The PNB-

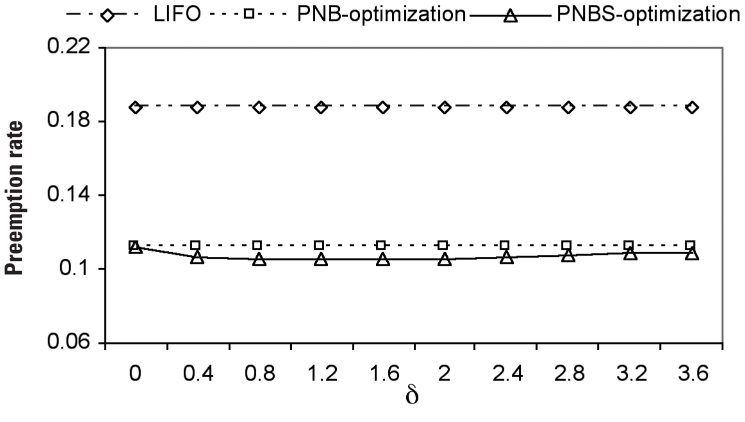


Figure 6: Preemption rate in different preemption policies

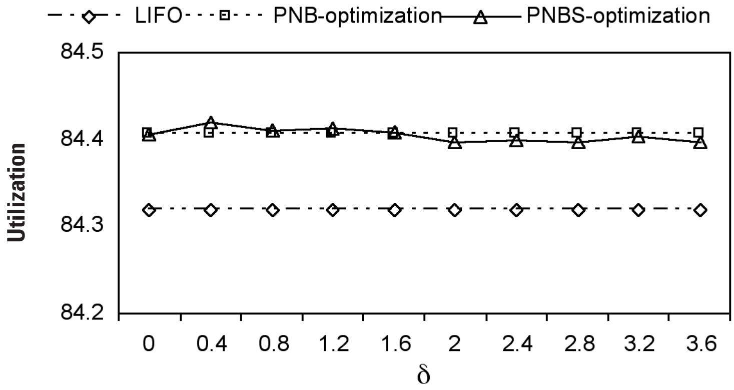


Figure 7: Link utilization in different preemption policies

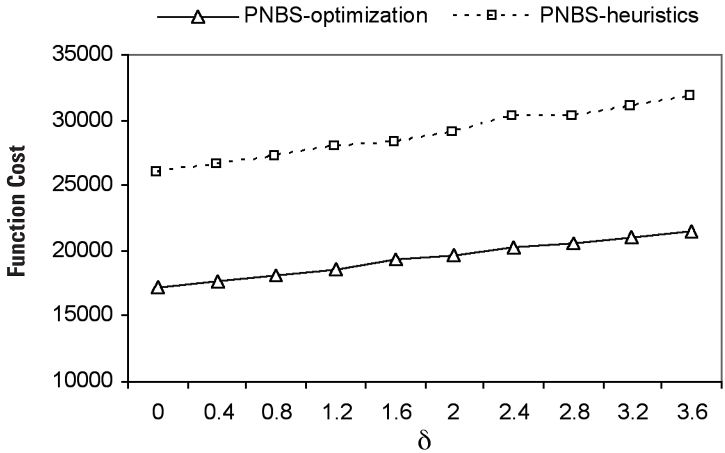


Figure 8: Preemption cost in optimization and heuristics

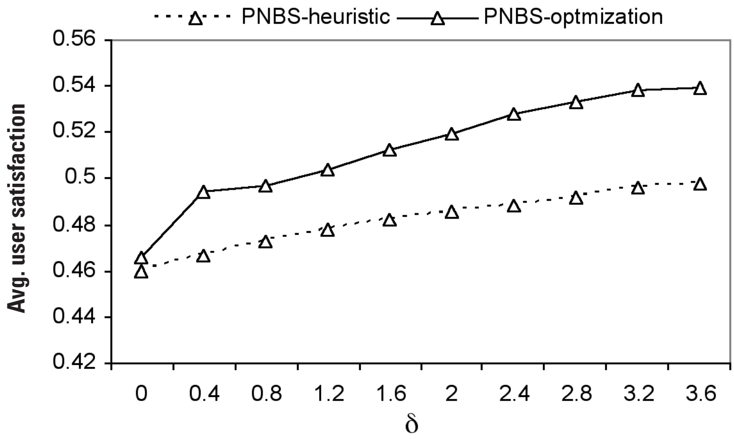


Figure 9: Average level of user satisfaction in optimization and heuristics

optimization policy tends to preempt more calls in order to minimize wasted bandwidth, the third criterion in the objective function. Observation suggests that PNB-optimization selects more calls to preempt with relatively smaller bandwidth to minimize the wasted bandwidth. This results in a slightly higher preemption rate in comparison to PNBS-optimization. PNB-optimization selects calls for preemption without any consideration of duration and it preempts many calls at their early lifetime. Conversely, PNBS-optimization considers the duration of a connection for calculation of estimated user satisfaction and tries to minimize the loss in user satisfaction. This is why PNBS-optimization is found to closely match PNB-optimization in terms of utilization (Figure 7) even after an addition of a new criterion. For $\delta <=1.6$, PNBS-optimization achieves better utilization and for $\delta >1.6$, PNB-optimization achieves slightly better utilization (<0.01%). From Figures 4 to 7, it can be concluded that the improvement achieved by the PNBS scheme in terms of average user satisfaction and priority of preempted calls does not result in detrimental affect on other network performance parameters like preemption rate and resource utilization.

Results obtained by using the corresponding heuristic are shown in Figures 8 and 9. Figures indicate that the heuristic follows the trend of optimization with increasing d. Further investigations suggest that the PNBS-heuristic outperforms the PNB-heuristic in terms of user satisfaction while both perform closely in terms of preemption rate and utilization.

6.2 Preemption Policy with Price Category

For simulation of the proposed PNBSP (priority, number of preempted calls, bandwidth and user satisfaction with price weight)-optimization (Eq. 9), we have considered three different categories of applications: i) real-time statistical bit rate applications, ii) deterministic bit rate applications, and iii) non real time statistical bit rate applications. Prices for these three different categories in $/Mbps are 0.2, 0.1 and 0.005 respectively (Morris and Pronk, 1999). For calculation of the RI index, user satisfaction levels need to be defined in four groups (Lewis, 2000) as mentioned in Section 5.3. For results shown in this section, we grouped users based on satisfaction as follows: $US_i = 1.0$: totally satisfied, 0.6~0.99: somewhat satisfied, 0.3~0.6: somewhat dissatisfied, 0~0.3: totally dissatisfied. Simulation was also conducted with groups based on other ranges of user satisfaction. The findings are consistent with the results reported in this section.

Figures 10 to 12 show the results at different preemption policies. High levels of user satisfaction are quite prominent in the PNBSP-optimization (Eq. 9) policy. PNBSP-optimization attains improvement by more than 2% over PNB-optimization and by 9% over LIFO preemption policy for an increasing value of δ (Figure 10). Figure 11 shows the particular improvement achieved by PNBSP-optimization with price information because priority of preempted calls is no longer as low as it is in Figure 5. This is because weighted dissatisfaction is considered in the objective function for PNBSP which ensures that calls with higher priority enjoy the privilege even when dissatisfaction is a consideration. Preemption rates in the PNB and PNBSP policies were found to match very closely to each other. For δ =0~3.6, the preemption rate in PNBSP-optimization varies in between 0.1123~0.1157 while for PNB optimization preemption rate remains at 0.1123. For link utilization PNBSP-optimization performs comparably to PNB-optimization. PNB-optimization achieves slightly better utilization for higher values of δ (>1.8). This is because of the relatively low contribution of the third criterion (wasted bandwidth in Eq. 9) to the objective function when the value of δ increases. However, decrease of utilization levels obtained in PNBSP-optimization is minimal (less than 0.01% at the worst case). Both PNBS-optimization and PNB-optimization achieved higher utilization in comparison with the LIFO approach as the LIFO approach does not use any information about bandwidth wastage.

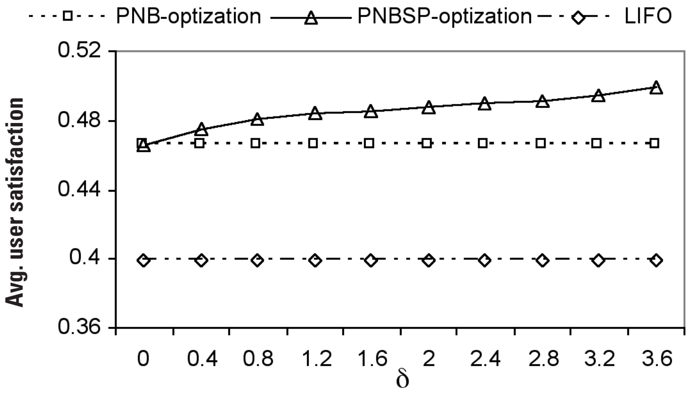


Figure 10: Average level of user satisfaction in different preemption policies with price category

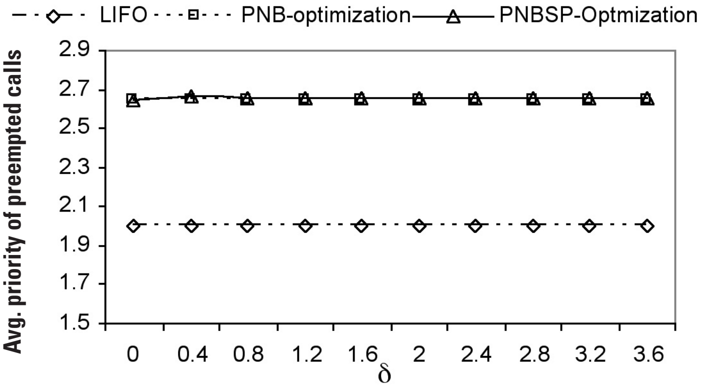


Figure 11: Average priority of preempted calls in different preemption policies with price category

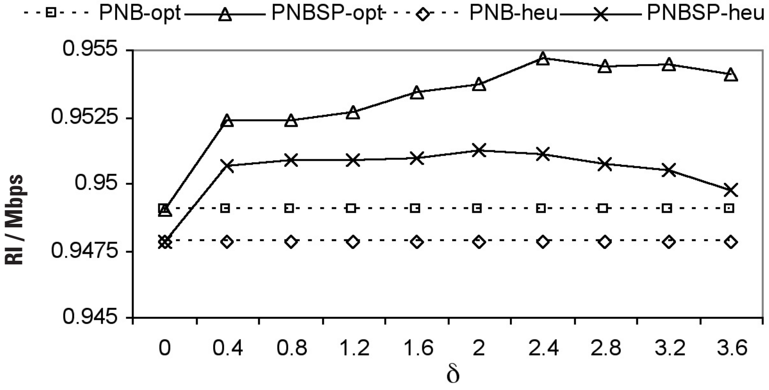


Figure 12: Revenue Index (RI) in different preemption policies

Results shown so far in this section indicate that PNBSP-optimization achieves better results in terms of user satisfaction for all values of δ and fit closely in preemption rate and utilization with PNB-optimization. Since neither PNB-optimization nor PNBSP-optimization achieve better results for all four criteria, it is not a straightforward decision for the network provider to select which one is best suited for a network. However, considering the economic aspect Revenue Index (RI) is certainly a useful single metric to show the trade-off between different criteria. Priority of preempted calls, preemption rate and user satisfaction have direct impact on the RI index. The product of RI per unit bandwidth and link utilization shows the impact of the four criteria on RI per unit time. Figure 12 shows the RI achieved by the PNB and PNBSP policies. The figure indicates that PNBSP-optimization achieves a much better revenue index in comparison with PNB-optimization. The LIFO approach (not shown in the figure) achieves the lowest RI (0.6879). This is because it preempts a higher number of calls having higher priorities and it does not consider the user satisfaction issue. Figure 12 shows that PNBSP-optimization achieves improved performance over the other in terms of RI for all values of δ (>0). At δ =2.4, RI yielded by PNBSP differs by 0.54% from PNB-optimization and 26.68% from LIFO approach. The relative decrease in utilization at δ =2.4 is less than 0.006% for PNBSP-optimization from PNB optimization. Considering total utilization, the revenue index/sec in PNBSP-optimization improves by about 0.54% from PNB-optimization. The revenue index found from the use of the heuristic incorporating the pricing information (Eq. 11) is also shown in Figure 12. It confirms that the improvement in the RI is evident independent of optimization and heuristics if the issue related to user satisfaction is taken into consideration.

6.3 Preemption Policy with Revenue Index

Further investigation was conducted aiming to minimize loss in revenue index instead of user dissatisfaction. User level satisfaction is mapped into the revenue index and then added to the objective function. Simulation results show that when the revenue index per call is added to the objective function and optimized, it achieves a higher revenue index (Figure 13). In the figure PNBSR denotes the policy considering the revenue index as the criterion in the objective function (Eq. 13). For most of the values of δ, PNBSR-optimization outperforms the other two preemption policies. For δ =2, the improvement in terms of RI is more than 0.18% over PNBSP-optimization and 0.67 % over PNB-optimization. However, for higher values of δ, PNBSP-optimization policy

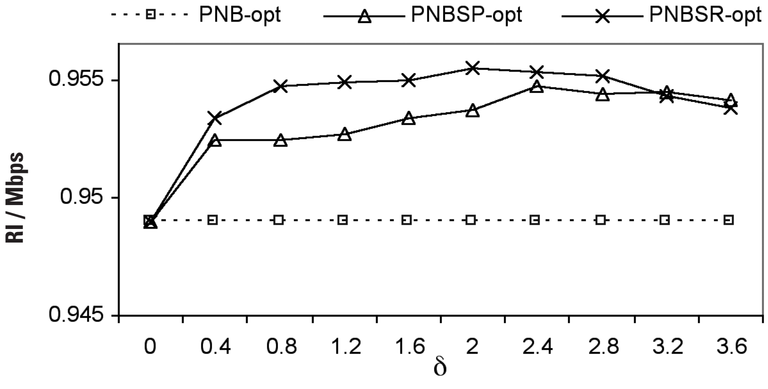


Figure 13: Revenue Index (RI) in different preemption policies

outperforms PNBSR-optimization policy in terms of revenue index. Although user satisfaction (Figure 14) is high in PNBSR-optimization, the combined impact of preemption rate and priority of preempted calls causes the drop in RI for PNBSR-optimization after a certain limit of δ.

In terms of utilization and preemption rate PNBSR-optimization closely follows the PNBSP and PNB-optimization policies. Results obtained from the heuristic (Eq. 15) are reported in Figure 15. These confirm that the PNBSR-heuristic performs better than the other two policies.

6.4 Impact of Co-efficients

The impacts of changing α, β, γ on user satisfaction and the revenue index were also investigated. Figure 16 shows the impact of changing α, the co-efficient associate with the preemption priority, (β=γ=δ=1) on the revenue index. It shows that increasing α has a clear impact on the revenue index. The revenue index increases in all policies as increasing values of α put more importance on the criteria of priority of preempted calls. For all values of α, the PNBSP and PNBSR policies ensure higher user satisfaction compare to PNB policy and consequently achieve better revenue index. The impact of changing β (α=γ=δ=1) is investigated in Figure 17. It shows that increasing values of β, the co-efficient associated with the number of preempted calls, achieves a decreasing RI per unit bandwidth. Further observation confirms that increasing β achieves improved preemption rate, but

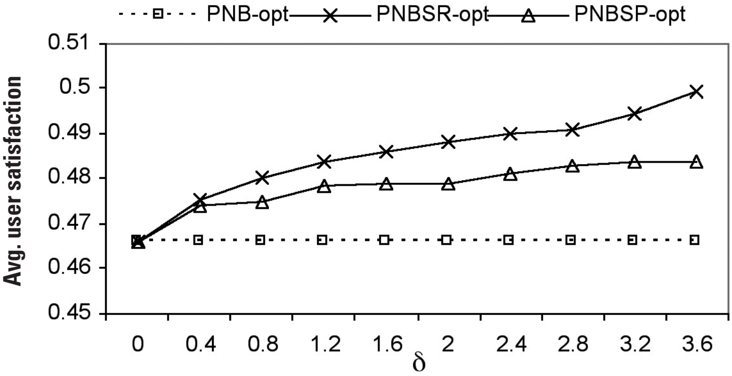


Figure 14: Average level of user satisfaction in different preemption policies

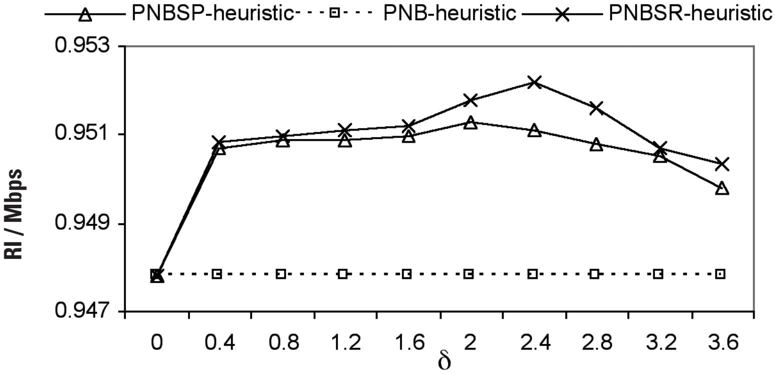


Figure 15: Revenue Index in different preemption policies (Heuristics)

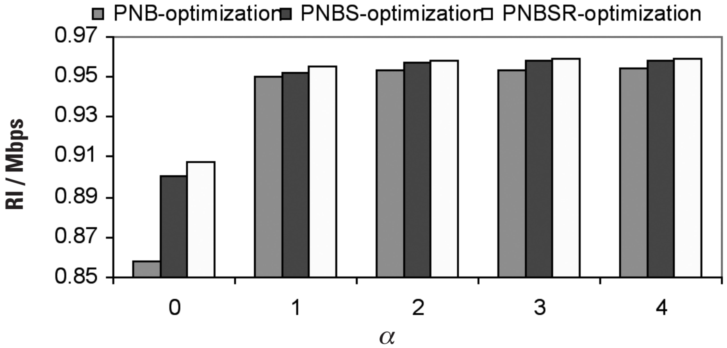


Figure 16: Revenue index for changing α

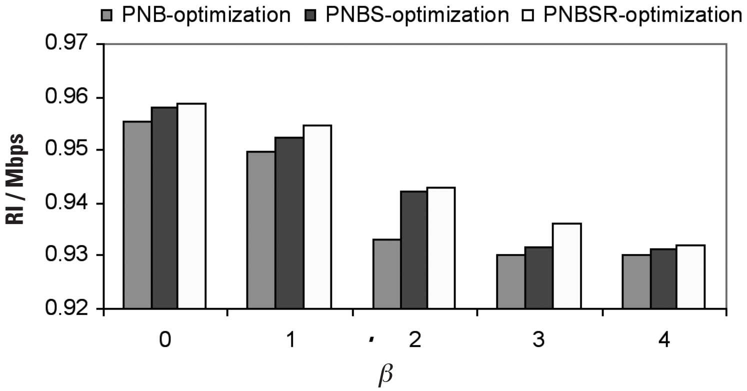


Figure 17: Revenue index for changing β

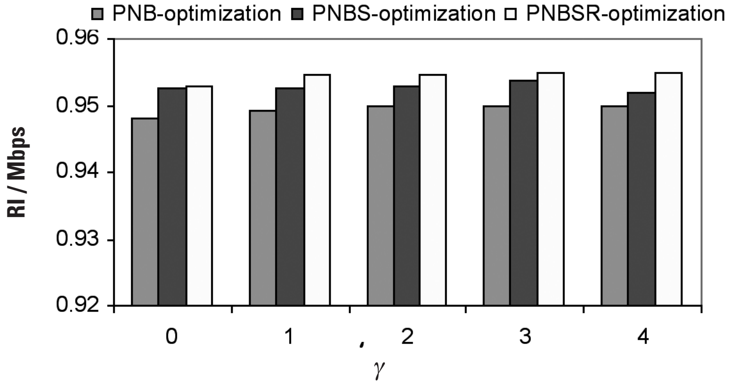


Figure 18: Revenue index for changing γ

since other co-efficients are kept the same, increasing β places relatively less importance on the priority level and user satisfaction. This brings detrimental affects on these criteria and decreases RI.

Figure 18 shows the RI for increasing values of γ. Results show that changes of γ, the efficient associated with wasted bandwidth, have small impact on RI per unit bandwidth. For all values of α,

β, γ PNBSR-optimization is found to provide the highest revenue index. Both PNSBP and PNSBR-optimization policies are found to outperform PNB-optimization and the LIFO policy consistently in terms of user satisfaction and revenue index. Although the simulation results presented in this section are shown for BA limit=0.8, simulations were also conducted for other BA limits within the range of 0.4 to 1.0 in a step of 0.1. Results at other BA limits also showed similar trends and the proposed policies consistently showed better performance compared to PNB and LIFO policy.

Categorization of user groups based on the value of user satisfaction $\mathrm{US_i}$ is required for the calculation of the revenue index. Results are shown in this section for one particular category. Further investigations were carried out to see how categorization of each user group in terms of $\mathrm{US_i}$ influences performance. Two other categorizations considered for the experiments are: category I: $\mathrm{US_i}$ =1.0: fully satisfied, 0.7~0.99: somewhat satisfied, 0.25~0.7: somewhat dissatisfied, 0~0.25: totally dissatisfied, category II: $\mathrm{US_i}$ =0.9~1.0: fully satisfied, 0.55~0.9: somewhat satisfied, 0.15~0.55: somewhat dissatisfied, 0~0.15: totally dissatisfied. Experiments with these categorizations of user groups consistently confirmed the improved performance achieved by PNBSP and PNBSR–optimization policies.

7. CONCLUSION

In this paper, we proposed and investigated new preemption policies for a QoS-enabled network. The key objective was to incorporate a new criterion, estimated customer satisfaction, in decision making for preemption. A user paying a higher price per unit bandwidth should be the less probable candidate for preemption. Following this argument, the proposed policy takes the valued user satisfaction into consideration in formulating the preemption policy. We also introduce another metric called the Revenue Index to measure the revenue earning prospect of the proposed policies. The merit of the preemption policies proposed in this paper lies in the fact that they achieve higher user perceived satisfaction and revenue index while maintaining very much the same level of resource utilization, priority of preempted calls and preemption rate achievable in the preemption policy proposed by Oliveira *et al* (2002). Among the three proposed policies, the preemption policy that directly incorporates information about the loss in revenue index per call basis was found to perform the best in terms of user satisfaction and the revenue index. As argued by Lewis (2002), higher revenue index indicates higher prospects of revenue return. In that respect the proposed policies will ensure higher revenue prospect and better user satisfaction for the network provider. Time and computational complexity were also considered and heuristics were presented for each optimization problem.

REFERENCES

AWERBUCH, B., AZAR, Y. and PLOTKIN, S. (1993): Throughput competitive online routing, *Proc. IEEE Symp. On Foundations of Computer Science*: 32–40.

AWDUCHE, D.O. (1999): Requirements for traffic engineering over MPLS, IETF, RFC 2701.

AHMAD, I., KAMRUZZAMAN, J. and ASWATHANARAYANIAH, S. (2004): Dynamic look-ahead time in Book-ahead reservation in QoS-enabled networks, *Proc. IEEE International Conference on Networks (ICON'04)*: 566–571.

AHMAD, I., KAMRUZZAMAN, J. and ASWATHANARAYANIAH, S. (2005): Improved preemption policy for higher user satisfaction, *Proc. 19th IEEE Int. Conf. on Advanced Information Networking and Applications (AINA) 2005*, 1: 749–754.

ANCLES (2005), http://www1.tlc.polito.it/ancles/, Accessed January 2005.

BRADEN, R., ZHANG, L., BERSON, S., HEROZ, S. and JAMIN, S., (1997): Resource reservation protocol (RSVP) – version 1 functional specification, RFC 2205, *Internet Engineering Task Force*.

CAMPANELLA, M., CHIVALIER, P. and SIMAR, N. (2001): Quality of Service Definition, http://www.dante.net/sequin/QoS-def-Apr01.pdf, 2001.

DO, M., PARK, Y. and LEE, J. (2002): Channel assignment with QoS guarantees for a multiclass multicode CDMA system, *IEEE Transactions on Vehicular Technology*, 51(5): 935–948.

FAUCHEUR, F. (2001): Requirements for support of Diff-Serv-Aware MPLS traffic Engineering, IETF Internet Draft.

GARAY, J. and GOPAL, I. (1992): Connection preemption in communication networks, *Proc. IEEE Infocom'92,* 1043–1050.

GREENBERG, A.G., SRIKANT, R. and WHITT, W. (1999): Resource sharing for book-ahead and instantaneous-request calls, *IEEE/ACM Trans. Networking*, 7: 10–22.

HARDY, W.C. (2001): QoS measurement and evaluation of telecommunications quality of service, John Wiley & Sons Ltd, New York.

IERA, A., MOLINARO, A. and MARANO, S. (2000): Call admission control and resource management issues for real-time VBR traffic in ATM-satellite networks, *IEEE Journal on Selected Areas in Communications* 18(11): 2393–2403.

LEWIS, S. (2002): Measuring the relationship between satisfaction and spending, articles in velocity 2002, http://development2.com/pdfs/velocity.pdf, Accessed January 2005.

LINDO API 2.0 (2005): www.lindo.com, Accessed January 2005.

LIN, Y., CHANG, C. and HSU, Y. (2002): Bandwidth brokers of instantaneous and book-ahead requests for differentiated services networks, *IEICE Trans. Commun.*, E85-B(1): 278–283.

MORRIS, D. and PRONK, V. (1999): Charging for ATM services, *Communication Magazine*, IEEE, 37(5): 133–139.

OLIVEIRA, J., SCOGLIO, C., AKYILDIZ, I. and UHL, G. (2002): A new preemption policy for diffserv-aware traffic engineering to minimize rerouting, *Proc. IEEE Infocom 2002*, 2: 695–704.

PNNI Draft specification (1995), ATM Forum 95-0471R14.

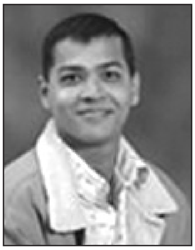
PEYRAVIAN, M. and KSHEMKALYANI, A. (1997): Connection preemption: issues, algorithms, and a simulation study, *Proc. IEEE Infocom* '97, 1: 143–151.

WANG, X. and SCHULZRINNE, H. (2000): Adaptive reservation: a new framework for multimedia adaptation, *Proc. ICME 2000*, 2(1): 1051–1054.

YI, D. and KIM, J. (2002): Dynamic resource management technique with advance reservation over QoS-provisioned networks, http://netmedia.kjist.ac.kr/jongwon/papers/2002pa-donghoon.pdf

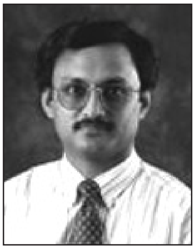
YAO, J., MARK, J., CHEW, Y., LYE, K. and CHUA, K. (2004): Virtual partitioning resource allocation for multiclass traffic in cellular systems with QoS constraints, *IEEE Transactions on Vehicular Technology*, 53(3): 847–864.

BIOGRAPHICAL NOTES



Iftekhar Ahmad

Iftekhar Ahmad received his B.Sc. in Computer Science and Engineering from Bangladesh University of Engineering & Technology (BUET), Dhaka, Bangladesh. He is currently doing his PhD in the Faculty of Information Technology, Monash University, Australia. Before joining Monash University, Iftekhar Ahmad worked as a Lecturer with IIUC, Bangladesh. His research interest includes computer networks, and computational intelligence.



Joarder Kamruzzaman

Joarder Kamruzzaman received his B.Sc. and M.Sc. in Electrical and Electronic Engineering from Bangladesh University of Engineering & Technology (BUET), Dhaka, Bangladesh and his PhD in Information System Engineering from Muroran Institute of Technology, Japan. He is currently a faculty member in the Faculty of Information Technology, Monash University, Australia. Before joining Monash University, Dr Kamruzzaman worked with James Cook University, Australia and BUET, Bangladesh. His research interest includes computer networks, computational intelligence and bioinformatics. He has published more than 90 peer-reviewed research publications including 27 journals. He has edited two reference books and served as a program committee member of a number of international conferences. Dr Kamruzzaman is a member of IEEE.