REVENUE AWARE PREEMPTION POLICY IN MULTIMEDIA COMMUNICATION NETWORKS

Iftekhar Ahmad, Joarder Kamruzzaman, and Srinivas Aswathanarayaniah
Gippsland School of Computing and Info Tech, Monash University, Victoria-3842, Australia

ABSTRACT

Preemption of low priority call connections to accommodate high priority call connections is a widely recommended technique in a multimedia communication network. A preemption policy is required to make decision about which connections to preempt when resource scarcity is experienced. In literature, priority of connection, preempted bandwidth and the number of preempted connections are proposed as the three basic criteria governing a preemption policy. So far revenue prospect of a communication provider, specially in relation to consumer satisfaction, has not been incorporated in preemption policy. In this paper we introduce revenue index, a metric that indicates the level of estimated consumer satisfaction, as an additional criterion to be used in conjunction with the above three. Revenue index is an indicative of long term revenue prospect of an enterprise. We present formulation of the preemption policy as an optimization problem. A heuristic to approximate the optimal solution is also derived. Simulation results show that the proposed model of preemption policy when adopted by a network provider ensures better consumer satisfaction for the end users which leads to higher revenue index for the network provider.

1. INTRODUCTION

The issue of bandwidth reservation and management has been attracting increasing interests from the researchers due to growing use of multimedia and distributed applications. Both multimedia and distributed applications require strict Quality of Service (QoS) which is guaranteed mainly by the implementation of adequate resource reservation in a QoS-enabled network. Since the amount of resources is limited, contention for resources among multiple call connections is a common scenario. Differentiation of call connection is required to solve the contention and ensure connection specific quality of service. One of the widely used techniques to ensure connection specific QoS is the preemption of less privileged calls to supply enough resources for high priority calls when resource scarcity is experienced. Preemption technique is governed by a preemption policy that determines which call to preempt under resource scarcity. For its high importance, formulating an optimal preemption policy has drawn the attention from researchers over a long period of time. Garay *et al.* [1] addressed the call preemption problem in a centralized network environment. They demonstrated that the process of selecting which calls to preempt in 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. A heuristic was presented to avoid such computational intractability. However, most of the resource reservation protocols proposed in the recent past like RSVP, RSVP-TE or ATM signalling use decentralized computation where each node has to make decisions and perform functions independent of other control points. Considering decentralized architecture, Peyravian *et al.* [2] proposed two algorithms: Min_Conn and Min_BW, which are computationally tractable to find the calls to be preempted in a decentralized architecture. Min_Conn algorithm first minimizes the number of call connections to be preempted, then searches the combination of connections to minimize the bandwidth to be preempted and if there is a choice of such combinations, it selects a combination that has the least priority levels of connections. Min_BW algorithm finds a solution in the order of importance of bandwidth, priority and number of connections.

Oliveira *et al.* [3] 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 desired criteria. The three criteria considered in the above work are: i) number of connections, ii) priority of connections to be preempted, and iii) amount of bandwidth to be preempted. However, service continuity of calls which is perceived by users as of outmost importance in a QoS-enabled network [4] 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. User dissatisfaction results in revenue loss. The objective of this paper is to introduce another optimization criterion termed as 'revenue index' modelled after consumer satisfaction in addition to the other three previously proposed optimization criteria in literature. Simulation results show that the proposed policy attains higher revenue prospects and better consumer satisfaction.

2. PROBLEM FORMULATION

Consider a connection request i, with bandwidth requirement b_i and priority p_i. If $\sum b_{j\in S} + b_i > C$, then find a set $U \in S$ such that $p_{j\in U} < p_i$, $\sum b_{j\in U} \geq \sum b_{j\in S} + b_i - C$ where S is the set of call connections currently using the link and C is the link

0-7803-9332-5/05/$20.00 ©2005 IEEE

capacity. All the elements of set *U* are with the attribute 'preemption enabled'.

Preemption policy finds the set *U* in response to the resource scarcity. The solution which closely fits with the problem statement is proposed by Oliviera *et al.* [3]. Mathematical formulation of this policy is given as

$$F(\mathbf{z}) = \alpha(\mathbf{z} \cdot \mathbf{y}^{\mathbf{T}}) + \beta(\mathbf{z} \cdot \mathbf{1}^{\mathbf{T}}) + \gamma(\mathbf{z} \cdot \mathbf{b}^{\mathbf{T}}) \quad (1)$$

The vector **z** is an optimization variable and is composed of *n* binary variables where *n* is the number of on-going preemption enabled call connections in the system.

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

$\mathbf{z}.\mathbf{y}^{\mathbf{T}}$ represents the priority of the preempted calls, $\mathbf{z}.\mathbf{1}^{\mathbf{T}}$ represents the number of preempted calls and $\mathbf{z}.\mathbf{b}^{\mathbf{T}}$ represents the total preempted bandwidth. $\alpha, \beta,$ and γ are the weights that select the level of preference. The solution of the problem stands as to minimize objective function $F(\mathbf{z})$ subject to the following constraint

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

A heuristic rather than an optimization solution is a better choice for a large size network in consideration of time complexity. Oliviera *et al.* [3] proposed a heuristic for a large size network as

$$H(l) = \alpha y(l) + \beta + \gamma (b(l) - r)^2 \quad (2)$$

$$y(l) = Q - \Pr(l)$$

where $y(l)$ indicates the loss in priority based on Q being the lowest priority level, $\Pr(l)$ being the priority of call connection *l* and lowest priority is denoted by highest numerical number. $b(l)$ indicates the bandwidth of the call connection *l*. H is calculated for each call and the calls are preempted in order of increasing value of H. In context of Book-Ahead (BA) reservation, Greenberg *et al.* [5] proposed to preempt the Instantaneous Request (IR) calls in order of Last In First Out (LIFO) fashion. The argument was if the call with the most recent arrival time was preempted the impact on successfully transmitted amount of data would be minimal.

3. PROPOSED PREEMPTION POLICY

Importance of service continuity on user satisfaction is different for different applications. For an application which performs an atomic task over its complete duration, utility gain is zero unless it is fully complete. An application whose importance increases sharpely towards the end of its completion like 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 chat have different satisfaction curve. We define the user satisfaction (gain) function as:

$$US_i = \begin{cases} e^{k\left(\frac{T_i}{D_i} - 1\right)} & \text{if } T_i < D_i \\ 1 & \text{if } T_i \geq D_i \end{cases} \quad (3)$$

where T_i is the time of data transmission before preemption of call connection *i* and D_i is the complete data transmission time of connection *i* if not preempted. However, only the end users have the idea of real D_i whereas the network provider can at best estimate it. In this paper, we model D_i equal to mean time of data transmission for calls of similar type (group) to which call connection *i* belongs. Value of *k* is application specific and it indicates the importance of user satisfaction on service continuity. It is important to mention that D_i can be different for different groups of applications and it can be obtained from distributions of applications in real time networks. US_i denotes the estimated level of user satisfaction when calculated on network provider side (to be used in preemption policy) whereas it shows the actual level of satisfaction when calculated on users' side (as reported in results section). Without loss of generality we assume that once a call is preempted then resumption of data transmission, if possible through rerouting is considered as a separate call connection.

Revenue return is one of the main driving forces for a network provider. In economics, the prospects of revenue are often determined by the level of user satisfaction. User satisfaction is a very important consideration for a network provider, specially when the long term future of the enterprise is considered. In a recent study, Lewis [6] proposed a measurement of the relationship between customer satisfaction and revenue prospects. A metric called as Revenue Index (RI) was proposed by Lewis that reflects the relationship between customer satisfaction and revenue return. A two step calculation of RI index was proposed as follows [6]:

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 calculation is based on the observation [6] 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 data collected over a long period of time. In this work we used the formulation proposed by Lewis to map estimated user satisfaction to Revenue Index.

In this paper we calculate and map estimated user satisfaction into RI and then use RI as one of the criteria for preemption policy. We propose to minimize weighted loss in RI in addition the other three criteria and the objective function is formulated as

$$F(\mathbf{z}) = \alpha(\mathbf{z}.\mathbf{y}^T)+\beta(\mathbf{z}.\mathbf{1}^T)+\gamma(\mathbf{z}.\mathbf{b}^T)+\delta(\mathbf{z}.\mathbf{x}^T) \quad (4)$$

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

$$x_i = \frac{p(i)}{\sum_{j=1 \text{ to } m} p_j}(1 - R_i)$$

where R_i indicates the level of revenue index for call connection *i* if it is placed for preemption, p_j is the price per unit bandwidth of 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 where *i* benongs. Heuristic for the above optimization problem is stated as

$$H(l) = \alpha\, y(l) + \beta + \gamma\,(b(l) - r)^2 + \delta\,(b(l) - r)^2\, x(l) \quad (5)$$

where $x(l) = \frac{p(l)}{\sum_{i=1,\cdots,m} p_i}(1 - R_l)$

For different price category applications overall RI shown in section 4 is calculated in the integrated form as

$$RI = \frac{p_1}{\sum_{i=1,\cdots,m} p_i} RI_1 + \frac{p_2}{\sum_{i=1,\cdots,m} p_i} RI_2 + \cdots\cdots + \frac{p_m}{\sum_{i=1,\cdots,m} p_i} RI_m \quad (6)$$

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

4. SIMULATION RESULTS

Simulation of the proposed policies has been done in context of Book-Ahead (BA) reservation for multimedia traffic. Book-ahead reservation requires guarantee of resource availability in advance. On the contrary, an Instantaneous Request (IR) call connection requires resource reservation instantly before activation. A BA call enjoys preference over an IR call because it books resources in advance. A preemption policy plays a very important role when a BA connection becomes active and requires resources to be preempted in a scenario where resources are shared between IR and BA call connections [7]. A single bottle-neck topology used for the simulation remains the same used in a number of related research works [7-8]. The capacity of each link is assumed to be 10 Mbps. IR call arrivals to the core link are assumed to follow Poisson distribution with a mean arrival rate of 11 calls per minute. Arrival of BA calls is also a Poisson distribution with a mean arrival interval of 50 sec. Bandwidth demand for IR calls is assumed to be exponentially distributed with a mean of 256 kbps. Bandwidth requirement of each BA call is exponentially distribution with a mean of 1.25 Mbps. To nullify the impact of difference in mean call holding time, 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 upto 80% of link capacity. Traffic analysis shown in this section is for the core link. Since a multiple bottleneck topology is basically a collection of multiple core links, traffic analysis of a single core link provides the results at the root level and thus works as the basis.

For simulation of proposed policy we have considered 3 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 are considered as 0.2, 0.1 and 0.005 dollar per unit bandwidth respectively [9]. For calculation of RI, user satisfaction level has to be defined in four groups as mentioned in section 3. For results shown in this section, we have defined the groups based on user satisfaction as follows: US_i=1.0: totally satisfied; 0.6~0.99: somewhat satisfied; 0.3~0.6: somewhat dissatisfied; and 0~0.3: totally dissatisfied. Simulation with other ranges of satisfaction level was also studied and it showed outcome consistent with the results reported in this section. User satisfaction and revenue index shown in this section is based on actual lifetime of a call (users' end) whereas the same in Eq. 3 and 4 calculated for implementation of preemption policy is based on predicted value of lifetime (network enterprise's end).

We investigated a number of important network parameters e.g., revenue index per unit bandwidth, average user satisfaction and bandwidth utilization. Simulation results show that when revenue index based on estimated user satisfaction per call is added to the objective function and optimized, it achieves higher revenue index at users' end. Figure 1 shows that the proposed policy 'PNBR' (priority, number, bandwidth and revenue as shown in Eq. 4) that considers revenue index as the additional criterion in objective function outperforms the PNB (priority, number, and bandwidth) preemption policy (Eq. 1) for most of the values of δ. For δ=2, the improvement in terms of RI is about 0.67% over PNB-optimization. The improvement of PNBR policy over LIFO policy was observed to be more than 25% for almost all the values of δ. Figure 2 shows the improvement in user satisfaction achieved in PNBR

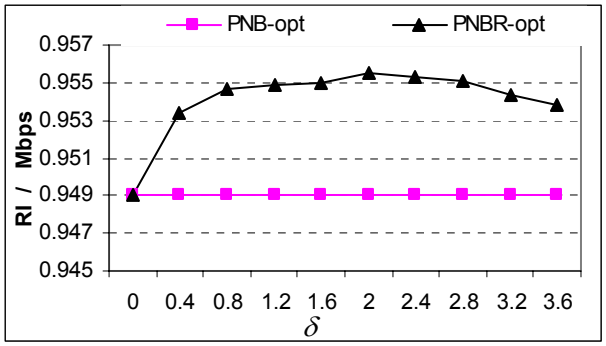

Figure 1: RI at different preemption policies.

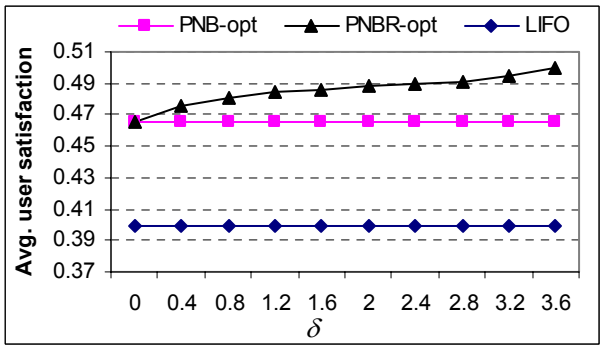

Figure 2: User satisfaction at different preemption policies.

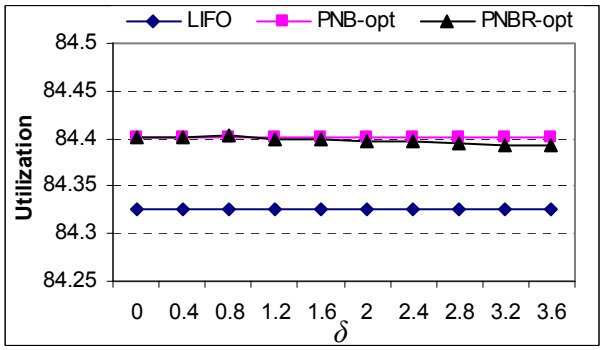

Figure 3: Utilization at different policies.

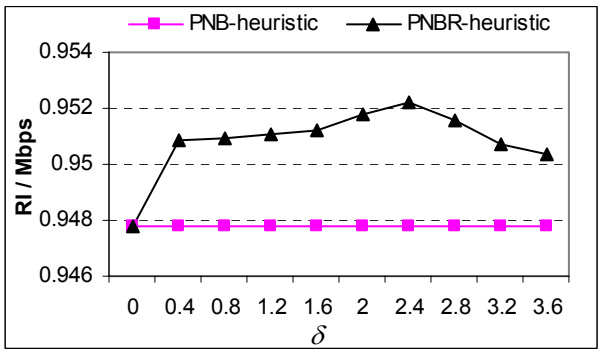

Figure 4: RI at different preemption policies using heuristic.

policy over PNB policy. It shows that increasing importance (δ) on the revenue index per call increases overall user satisfaction (>3% at $\delta \geq 3.2$). However, other considerations like priority and number of preempted calls which are emphasised less with increasing δ, also have impact on overall RI (Eq. 6) and this is why RI/Mbps shows downtrend after a certain value of δ (Fig.1). Figure 3 shows the utilization in different policies. The difference in utilization between PNB and PNBR policy is marginal. Result obtained from heuristic (Eq. 5) is reported in Fig. 4. It confirms that PNBR-heuristic also outperforms PNB-heuristic (Eq. 2).

5. CONCLUSION

In this paper, we proposed and investigated a new preemption policy for a multimedia communication network. The key aim was to formulate a new criterion termed Revenue Index modelled after consumer satisfaction and incorporate this criterion in preemption policy by formulating an objective function in an optimization problem. Simulation results show that the proposed policy incorporating estimated revenue index criterion outperforms existing preemption policies in terms of customer satisfaction and revenue index, and performs comparably in terms of resource utilization. Time and computational complexity were also considered and heuristic was presented for the optimization problem. As argued in [6], higher revenue index indicates higher prospects of revenue return. In that respect the proposed policy will ensure higher revenue prospect and better user satisfaction for the network provider.

6. REFERENCES

[1] J. Garay, and I. Gopal, "Connection preemption in communication networks," *Proc.* of *IEEE INFOCOM*' 92, pp. 1043-1050, 1992.

[2] M. Peyravian, and A. Kshemkalyani, "Connection preemption: issues, algorithms, and a simulation study," *Proc. IEEE INFOCOM '97*, vol. 1 , pp. 143-151, April,1997.

[3] J. Oliveira, C. Scoglio, I. Akyildiz, and G. Uhl, "A new preemption policy for diffserv-aware traffic engineering to minimize rerouting," *Proc. IEEE INFOCOM,* pp.695-704, 2002.

[4] M. Campanella *et. al.*, "Quality of Service Definition", http://www.dante.net/sequin/QoS-def-Apr01.pdf.

[5] A.G. Greenberg, R.Srikant, and W. Whitt, "Resource sharing forbook-ahead and instantaneous-request calls," *IEEE/ACM Trans. Networking*, vol.7, pp.10-22, Feb. 1999.

[6] S. Lewis, "Measuring the relationship between satisfaction and spending," articles in velocity 2002, http://development2.com/pdfs/velocity.pdf.

[7] I. Ahmad, J. Kamruzzaman, and S. Ashwathanarayaniah, "Dynamic look-ahead time in book-ahead reservation," *Proc. IEEE ICON '04*, pp. 566-571, Singapore 2004.

[8] Y.S. Sun, Y. Tu, and M.C. Chen, "Admission control and capacity management for advance reservations with uncertain service duration," *Proc. LNCS*, vol. 2345, pp. 190-201, 2002.

[9] D. Morris and V. Pronk, "Charging for ATM services," *IEEE Comm. Magazine,* vol. 37, no. 5, pp. 133-139, 1999.