DiffServ Node with Join Minimum Cost Queue Policy: Analysis with Multiclass Traffic

D. Manjunath, Dept. of Electrical Engg
Ashish Goel, Dept. of Electrical Engg
N. Hemachandra, IE & OR Programme

Indian Institute of Technology Bombay
Powai Mumbai 400076 INDIA
{dmanju@,ashgo@ee,nh@}iitb.ac.in

Abstract— **The DiffServ model is an attractive candidate for providing relative QoS in the Internet. This is also easily amenable to simple and effective pricing mechanisms. By pricing access to a relative QoS, we can model a DiffServ node as a "Join Minimum Cost Queue" in which an arriving customer (packet or connection) determines the relative cost as a function of the congestion in the different queues and their access prices and decides to take service from that queue for which the cost is minimum. The Paris Metro pricing scheme and its work conserving variant called the Tirupati pricing are analyzed in this paper in the presence of multiclass traffic and for static pricing. Two of the more interesting observations are that the disutility and revenue rate are not monotonic or convex functions of price and the revenue rate is very sensitive to the behavior of the delay sensitive class.**

Keywords— **DiffServ, Network Pricing, Paris Metro Pricing, Join Minimum Cost Queues.**

I. Introduction

A number of pricing models to save the Internet from suffering the "tragedy of commons" have been proposed and continue to be proposed. Considering the volume (measured in bytes, packets or connections) of traffic handled by the Internet, it is now well established that these schemes have to be simple to implement and be robust. Varaiya and Walrand [1] identify four categories of Internet charges – access, usage, congestion and quality. It is also generally believed that in this classification, it is computationally expensive to implement pricing structures for all but the access charges. See [2] for an excellent overview and bibliography on Internet pricing issues.

It is now believed that absolute end-to-end quality assurances is very difficult and probably infeasible in the Internet and relative guarantees on a hop-by-hop basis are more feasible. This has led to the differentiated services [3] emerging as the potential vehicle for QoS in the Internet. The differentiated service model divides the available bandwidth among the multiple classes of traffic in a static or dynamic manner. Network nodes maintain separate logical queues for each class for each outgoing link and service them according to the bandwidth sharing policy for that link. The simplest way to provide multiclass service is through priority queues but that could lead to starvation of lower priority queues for extended periods of time. A second option would be to have multiple queues and force a certain arrival profile to each queue to provide different QoS guarantees to arrivals to that queue. Pricing seems to be a good vehicle to provide this control. Paris Metro Pricing scheme (PMP) is one such pricing scheme for multiclass queues [4]. In the PMP scheme, the network is logically partitioned in a static manner, each partition is allocated a fixed share of the network resources and the access price for each partition is differentially priced. A similar model is proposed in [5] and is called Olympic pricing scheme. In [6], [7] the behavior of PMP under equilibrium conditions is considered and compared with a uniclass pricing scheme.

In [8] a work conserving version of PMP called the *Tirupati scheme*, was proposed. This scheme takes inspiration from the queue management scheme at Tirupati, a major pilgrimage center in southern India where it has been operating with remarkable efficiency for quite some time now. In [8], the social optimality of the Tirupati pricing model was analyzed and it was shown that the difference between the social cost of the optimally priced system and that of the Tirupati system is $\bar{K}\epsilon$ for constants $\bar{K}$ and ϵ. Another, more practical, contribution of [8] was dynamic pricing using a dynamic programming equation and a reinforcement learning based online pricing algorithm.

We consider pricing quality through an access charge and analyze pricing in a multiclass service system like the DiffServ model of the IETF [3], where the access charge is a function of the service class. We abstract an Internet node as a queueing system and analyze the Tirupati and PMP schemes under static pricing in the presence of multiclass traffic into this queue. Our results are from classical Markovian queuing analysis similar to the analysis of join shortest queue systems from where we borrow some techniques. The paper is organized as follows. In the next section we describe the models that we study and the assumptions that we will make. In Section III we describe the analysis method to obtain the bounds on the revenue rates and the approximate results on the disutility rates (to be defined later) in an infinite buffer system. In Section IV, we present the numerical results from our analyses of the Tirupati and PMP pricing models. Finally, we conclude in Section V.

II. Join Minimum Cost Queue with Multiclass Customers

The "Join Minimum Cost Queue (JMCQ)" that we consider works as follows. A single exponential server of capacity μ serves $K > 1$ queues. The K queues have different join prices $p_1 \leq p_2 \leq \cdots \leq p_K$. We assume Poisson arrivals at rate λ. Arrivals belong to M, $M \geq 1$ classes, and the class of an arriving customer is independent of others and chosen according to a fixed distribution $r = [r_1, \cdots, r_M]$, r_m being the probability of a customer being class m. Let $N_i(t)$ be the number of customers in queue i, $1 \leq i \leq K$ and let $\mathbf{N}(t) = [N_1(t), N_2(t), \ldots, N_K(t)]^T$ be the queue length vector at time t. A class m arrival at time t is informed of $\mathbf{N}(t^-)$, the queue vector 'just before t'. It calculates its disutility for queue i according to a function $\psi(m, N_i(t), p_i)$ and joins queue i if i minimizes $\psi(m, N_j(t^-), p_j)$ among all j. Ties in disutility are awarded to the lower priced queue. The disutility function $\psi(\cdot)$ should reflect the sensitivity to the price and the delay. This policy is clearly individually optimal and in [8] it was shown that for a linear disutility function the balanced JMCQ queue is also socially optimal for a single class traffic in the sense that the difference between the social cost of the optimally priced system and that of this system is $\bar{K}\epsilon$ for a constant ϵ.

The PMP uses a static partition of server capacity among the K queues - queue i is served at rate μ_i, $\sum_{i=1}^{K} \mu_i = \mu$. The Tirupati service system [8] is a work conserving version of PMP and in its most general form, it uses the generalized processor sharing model to service the K queues where queue i is allocated a weight w_i and receives service at rate μ_i. Here

$$\mu_i = \frac{w_i \mu}{\sum_{j:\text{ Queue } j \text{ is non empty}} w_j} \tag{1}$$

The JMCQ is balanced if the capacity is divided equally among all the queues. Thus the Tirupati queue is balanced if $w_i = w$ for all i, (the service rate for each queue will be $\mu/j, j \in [1, \cdots, K]$ where j is the number of nonempty queues) and the PMP is balanced if $\mu_i = \frac{\mu}{M}$ (the service rate for queue i will be μ_i irrespective of the occupancy of the other queues).

We prescribe only a join price and no price for the quantum of service. This is in the spirit of the Internet in having only a flat "access" charge and no volume charge and is simple to implement. The congestion level in terms of the queue length of each of the queues is posted to enable the customers to calculate their disutility. The mathematical abstraction that we consider is a queue with customer arrivals. Hence, we assume that the customers obtain instantaneous congestion information on arrival and that there is no delay as might be expected in the context of a an Internet. An important contribution of this paper is that we explicitly involve a model of customer behavior based upon a "disutility" calculated from the posted prices and congestion levels. We can also consider balking customers by having a type m customer balk from joining queue i if its disutility is strictly greater than some constant α_{im}. If this happens for all i, the customer leaves without joining any queue. In addition, we could also consider a system with finite buffers, queue i has room for at most B_i customers at any time. Thus, the balking mechanism can either be customer induced, corresponding to a sensitivity to the disutility (price and/or congestion) of the customer, or system induced, reflecting implicit prioritization by the system.

Let p_{im} be the join cost of queue i for class m customer. A convex combination of the queue length (the congestion indicator) and price as a disutility function is simple and effective in capturing price and delay sensitivities for different types of traffic. Thus the disutility function that we consider is

$$\psi(m, N_i(t), p_{im}) = a_{im} N_i(t) + (1 - a_{im}) p_{im} \quad a_{im} \in (0, 1) \tag{2}$$

The a_{im}, called the delay sensitivity, serves to scale the queue vector with respect to the price vector for each queue and also allows for differentiated scaling for the K queues among the different classes. In our analyses we assume that the service rate for all classes are identical and hence the prices are not class dependent, a reasonable assumption in the context of the Internet. Therefore, $a_{im} = a_m$ for $m = 1, \cdots M$ and $p_{im} = p_i$ for $i = 1, \cdots K$ and Eqn. 2 reduces to

$$\psi(m, N_i(t), p_i) = a_m N_i(t) + (1 - a_m) p_i \tag{3}$$

In the rest of the paper we will consider a balanced JMCQ with two customer classes and two queues, i.e., $M = 2$ and $K = 2$. We will also assume that $\mu_1 = \mu_2 = \mu/2$. Without loss of generality, we assume $a_1 > a_2$, i.e., class 1 traffic is delay sensitive and class 2 traffic is price sensitive. In the next section we analyze the system with infinite buffers and no balking.

III. Analysis of Infinite Buffer JMCQ

The JMCQ policy is clearly similar to the "Join Shortest Queue" (JSQ) policy that has been extensively studied. See, for example, [9] for some recent results. The exact performance analysis of the JSQ system has been notoriously difficult and only approximations and asymptotic results are available except for some very special cases. [10] gives a detailed survey of the analytical models for JSQ policy. The JMCQ considered here is further complicated by the existence of a cost associated with the queues and also the presence of multiple classes of customers. Thus it would be reasonable to assume that the exact closed form expressions for performance analysis will be difficult. Therefore, we consider an approximate analysis. We will concentrate on two kinds of results – the revenue rate for the server and the disutility rate for the customers.

The evolution of $\mathbf{N}(t)$ is modeled as a two dimensional birth death process. Given $\mathbf{N}(t)$, the departure process is governed by the service policy - non work conserving PMP or work conserving Tirupati. The queue into which an arrival that sees $\mathbf{N}(t)$ will join is determined by its class and its delay sensitivity. Define δ_{ij}^m to be the queue that an arriving class m customer to state $[i, j]$ will join. For example, $\delta_{ij}^1 = 1$ if $\psi(1, i, p_1) \leq \psi(1, j, p_2)$ and $\delta_{ij}^1 = 2$ otherwise. The transition rate matrix $\mathbf{Q} = \{q_{ij:kl}\}$, $i, j, k, l = 0, 1, \cdots$ is easily determined.

We assume that the Markov chain described by $\mathbf{Q}$ above is ergodic and a steady state solution $\underline{\pi} = \{\pi_{ij}\}$ exists. When the system is in state $[i, j]$, a revenue p_i is earned if an arrival joins queue i. Therefore, from the Law of Large Numbers, the steady state rate of revenue generation, $\mathcal{R}$, is given by [11]

$$\begin{aligned} \mathcal{R} &= \sum_{i,j} \pi_{ij} \lambda \left[r_1 p_1 I(\delta_{ij}^1 = 1) + r_1 p_2 I(\delta_{ij}^1 = 2) + \right. \\ & \left. + r_2 p_1 I(\delta_{ij}^2 = 1) + r_2 p_2 I(\delta_{ij}^2 = 2) \right] \end{aligned} \tag{4}$$

where $I(\cdot)$ is the indicator function with the usual meaning of taking on a value 1 if the condition is satisfied. Similarly, the disutility rate for class m customers, $\mathcal{D}_m$ is given by

$$\begin{aligned} \mathcal{D}_m &= \sum_{i,j} \pi_{ij} \lambda r_m \left[I(\delta_{ij}^m = 1)(a_m p_1 + (1 - a_m) i) + \right. \\ & \left. + I(\delta_{ij}^m = 2)(a_m p_2 + (1 - a_m) j) \right] \end{aligned} \tag{5}$$

We now describe the method to obtain bounds and approximate results. First consider the revenue rate, $\mathcal{R}$. We borrow

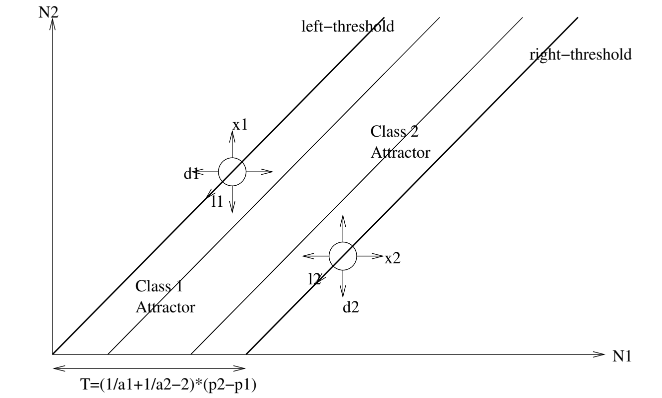

Fig. 1. Attractors for the two classes of customers with $a_1 > a_2$ are shown as thin lines. Truncation thresholds are shown as thick lines. Also shown are two example states on the thresholds and the transitions from them. JMCQ policy will not allow transitions marked x. Only transitions marked d_1 and d_2 will make the new system "leave" $\mathcal{S}$. For $JMCQ_u$, transitions d_1 and d_2 are dropped. For $JMCQ_l$, transitions l_1 and l_2 are introduced with the same rate as d_1 and d_2 respectively.

the techniques of [12] to obtain bounds on $\mathcal{R}$ by considering a truncated state space. To motivate the truncation, first consider the situation when only one class of customers is present. Without loss of generality, let $p_1 < p_2$. On the $N_1(t)$–$N_2(t)$ plane, an *attractor* line can be defined such that an arrival to the system when the state is on the left of this line will join queue-1 and an arrival to a state on the right of this line will join queue-2. This means that arrivals tend to move the system towards the attractor. The attractor line is defined by the equation $N_2 = N_1 - (\frac{1-a_m}{a_m})(p_2 - p_1)$, i.e., arrivals will prefer the lower priced queue unless $N_1 - N_2 > (\frac{1-a_m}{a_m})(p_2 - p_1)$. With multiclass traffic, there will be separate attractors for each class. Therefore, for states (i, j) away from the attractor we argue that π_{ij} becomes very small and we can consider a state space truncated at some distance from the attractors and suitably modified transition rates. Specifically, we will consider the state space of the JMCQ truncated to contain only those states for which $0 \leq |N_1 - N_2| < T, T \geq (\frac{1-a_2}{a_2} + \frac{1-a_1}{a_1})(p_2 - p_1)$. T will be called the truncation threshold. Let $\mathcal{S}$ denote the state space of the original JMCQ system and $\mathcal{S}'$ the state space obtained after truncation.

We define systems $JMCQ_u$ and $JMCQ_l$ over $\mathcal{S}'$, like in [12]. These give us upper and lower bounds on the mean delay and the mean number in the system for the JMCQ. First consider system $JMCQ_u$ and let $\mathbf{Q}_u = \{q^{(u)}_{ij:kl}\}$ be its transition rate matrix which is obtained from $\mathbf{Q}$ as follows. The transition rates from the states that are between the two truncation threshold lines will be the same as that in the JMCQ system. In the states on the truncation threshold line, an arrival will move the state space towards its attractor and will not cause the system to go out of $\mathcal{S}'$. In a state on the threshold line, a departure is disallowed. This means,

$$q^{(u)}_{ij:kl} = \begin{cases} q_{ij:kl} & \text{if } ij \text{ and } kl \text{ are states between threshold lines} \\ 0 & \text{if } ij \text{ is on left threshold; } k = i-1, j = l \\ 0 & \text{if } ij \text{ is on right threshold; } k = i, l = j-1 \end{cases}$$

Now consider system $JMCQ_l$ and its transition rate matrix $\mathbf{Q}_l = \{q^{(l)}_{ij:kl}\}$ which is obtained from $\mathbf{Q}$ as follows. As with $JMCQ_u$, the transition rates from the states between the threshold lines are the same as that in the JMCQ system. Once again, as with the $JMCQ_u$, we need to consider only the transition rates corresponding to departures from states on the truncation threshold line. These are modified such that a departure will take away a customer from the other queue. This means

$$q^{(l)}_{ij:kl} = \begin{cases} q_{ij:kl} & \text{if } ij \text{ and } kl \text{ are interior states} \\ q_{ij:i-1\,j} & \text{if } ij \text{ is on left threshold; } k = i-1, l = j-1 \\ q_{ij:i\,j-1} & \text{if } ij \text{ is on right threshold; } k = i-1, l = j-1 \end{cases}$$

Let $\overline{D}_{mu}$ the mean delay for class-m traffic and $\overline{N}_{iu}$ the mean number in queue i in the $JMCQ_u$ system. Also let $\overline{D}_{ml}$ and $\overline{N}_{il}$ be the corresponding quantities for the $JMCQ_l$ system.

Proposition 1: 1. $\overline{D}_{ml} \leq \overline{D}_m \leq \overline{D}_{mu}$
2. $\overline{N}_{il} \leq \overline{N}_i \leq \overline{N}_{iu}$

The proof is exactly similar to that in [12].

Let $\mathcal{S}'_{Tl}$ be the states on the left threshold line and $\mathcal{S}'_{Tr}$ the states on the right threshold line. Let π^u_{ij} be the stationary distribution of $JMCQ_u$ and define $\mathcal{R}_u$ as

$$\begin{aligned} \mathcal{R}_u \quad = \quad & \sum_{ij \in \mathcal{S}'} \left\{ \pi^u_{ij} \lambda \left[r_1 p_1 I(\delta^1_{ij} = 1) + r_1 p_2 I(\delta^1_{ij} = 2) \right.\right. \\ & \left.\left. + r_2 p_1 I(\delta^2_{ij} = 1) + r_2 p_2 I(\delta^2_{ij} = 2) \right] \right\} \\ & + \sum_{ij \in (\mathcal{S}'_{Tl} - 00)} \pi^u_{ij} \frac{\mu}{2} p_1 + \sum_{ij \in (\mathcal{S}'_{T_r} - T0)} \pi^u_{ij} \frac{\mu}{2} p_2 \quad (6) \end{aligned}$$

Along similar lines we can define $\mathcal{R}_l$.

Proposition 2: $\mathcal{R}_u \geq \mathcal{R} \geq \mathcal{R}_l$

Proof: We prove the first half of the above inequality. The $JMCQ_u$ and $JMCQ_l$ behave identically to the $JMCQ$ for states between the two threshold lines. Further, when the state is on a threshold line, the $JMCQ_u$ and $JMCQ_l$ behave identically to $JMCQ$ for arrivals and we need to consider only departures corresponding to transitions d_1 and d_2 of Figure 1.

- Disallowing transition d_1 increases the disutility of the lower priced queue leading subsequent arrivals to join the higher priced queue with a higher probability, hence increasing the revenue on an average. While the system is on the left threshold line, revenue p_1 is accumulated into $\mathcal{R}_u$ at rate $\frac{\mu}{2}$ in Eqn. 6.
- Similarly, d_2 is disallowed in $JMCQ_u$ and from Eqn. 6, $\mathcal{R}_u$ accumulates revenue p_2 at rate $\frac{\mu}{2}$ while the system is on the right threshold line. This more than compensates for the loss in revenue due to increase in disutility of queue 2.

The $JMCQ_l$ and $JMCQ_u$ systems can be solved as follows. Rename the states in $\mathcal{S}'$ as follows: (N_1, N_2) is renamed as $(N_1 - N_2, N_2)$ for $N_1 = N_2, \cdots N_2 + T$. Since the system is a two-dimensional birth death process, the non zero transition rates are only between "adjacent" states and the transition rate matrix for the $JMCQ_l$ and $JMCQ_u$ systems will have the block tridiagonal structure of a quasi birth death process and the matrix geometric techniques [13] are directly applicable. For more details of this mapping see [12].

IV. Numerical Results

We present some numerical results of our analysis of the PMP and Tirupati variations of the JMCQ and analyze the dependence

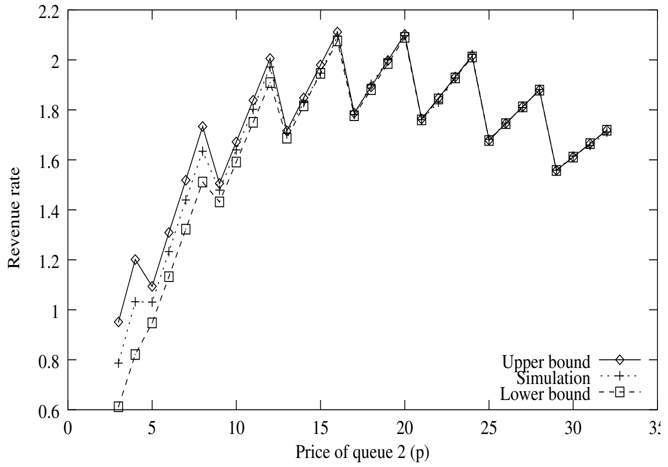

Fig. 2. Comparing the revenue rates from $JMCQ_u$ and $JMCQ_l$ against results from simulation. The truncation threshold is chosen at two levels to the right of the right attractor line, specifically, $T = \lceil \frac{7}{3}p \rceil + 2$. We compute using 100 stages of the QBD process. Other parameters are $\lambda = 0.8$, $a_1 = 0.8$ and $a_2 = 0.3$.

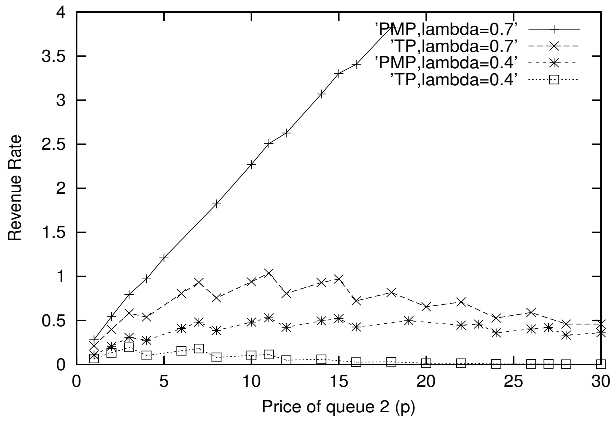

Fig. 3. Comparing revenue rates in the PMP and Tirupati pricing schemes for values of λ less than and greater than the service rate of each server of the PMP; $a_1 = 0.8, a_2 = 0.3$

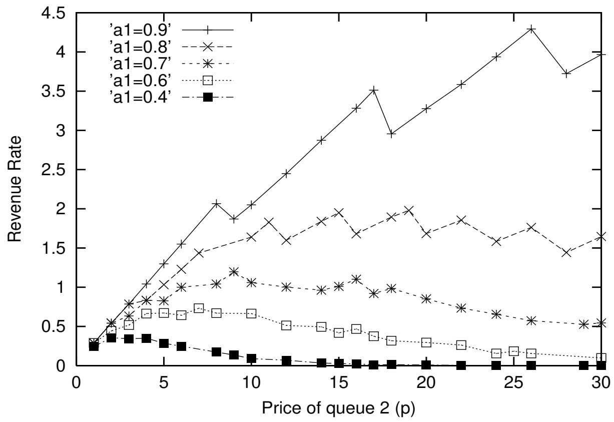

Fig. 4. Revenue rate as a function of p for different a_1; $\lambda = 0.8, a_2 = 0.3$.

of revenue and disutilities on the delay sensitivities of the customer types and the price vector. We will use $p_1 = 0$ and $p_2 = p$. Thus price variation will essentially mean a variation of p. Also, unless explicitly mentioned, $\lambda_1 = \lambda_2 = 0.5\lambda$, the arrivals are equally likely to belong to either traffic class. Further, the results are primarily for the Tirupati case unless explicitly mentioned.

First, consider the tightness of the bounds on the revenue rate. Figure 2, shows the upper and lower bounds on the $\mathcal{R}$ for $T = 4$ and $T = 15$ from analysis and that obtained from simulation. Observe that even with $T = 4$, the bounds are just 5–10% off of the simulation results. The bounds get tighter to within 1% for $T = 15$. Also observe that although the revenue rate generally increases with increasing p reaches a "maximum" and then decreases, it is neither monotonic nor a convex function of p. This means that $\mathcal{R}$ does not have a unique maxima.

Note that we have not established that the disutility rates obtained from either the $JMCQ_u$ or the $JMCQ_l$ systems are bounds on the actual disutility. We will use the $JMCQ_u$ system to obtain the disutility rates and treat the results as an approximation of the actual value rather than as an upper bound.

Next we compare the revenue rate in the Tirupati and PMP pricing models. In PMP, if $\lambda > \mu_1$, queue 1 will never empty and the system will always operate near the equilibrium lines. Increasing p will only move the equilibrium lines and $\mathcal{R}$ will be an increasing function of p. This is shown in Figure 3 for $\lambda = 0.7$. However, when $\lambda \leq \mu_1$, i.e., $\lambda = 0.4$, the revenue rate function of PMP looks similar to that of the Tirupati.

We now consider the effect of price sensitivity on $\mathcal{R}$ and $\mathcal{D}_m$ in the Tirupati queue. Figures 4 and 5 show the revenue rate as a function of the delay sensitivities of the two classes of traffic. Although $\mathcal{R}$ increases with increasing a_m, the following observations are interesting.

1. The change in the revenue rate due to change in the a_1 is much more than that caused by a change a_2. For example, even doubling of a_2 from 0.3 to 0.6 does not change the revenue rate appreciably.
2. The price at which the maximum revenue rate is achieved is very nearly the same for the different values a_2 whereas it increases for increasing values of a_1.

This is because a large fraction of the revenue is derived from the delay sensitive class and any changes in its behavior have a more noticeable effect than those in the price sensitive class. We also observed that the revenue is more sensitive to the delay sensitive class even when $(a_1 - a_2)$ is small.

Figure 6 shows the disutility rate as a function of p for different combinations of a_1 and a_2 for the Tirupati and PMP schemes. Increasing a_1 decreases the disutility rate while increasing a_2 increases it. Further, it is an increasing function of p. However, as the price increases, the disutility rate does not increase beyond a certain value. This is because all arrivals join queue 1 and therefore the disutility is less and less a function of p. For PMP, the disutility increases rather rapidly and linearly.

To study revenue rate as a function of the arrival rate, we consider two cases – (i) λ is varied $\lambda_1 = \lambda_2 = 0.5\lambda$, (ii) λ remains constant and λ_1 and λ_2 are varied. Figure 7 shows that the revenue rate behaves much like the mean delay function as a function of λ. This is expected because at high arrival rates, queue lengths can be significant and the delay sensitive traffic will choose the higher priced queue, thereby bringing in revenue. Figure 8 shows the revenue variation due to a change in the mix of the traffic classes. Once again, the revenue rate is

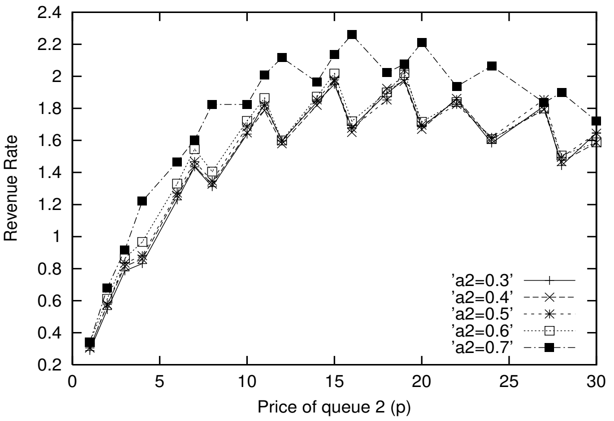

Fig. 5. Revenue rate as a function of p for different a_2; $\lambda = 0.8$, $a_1 = 0.8$.

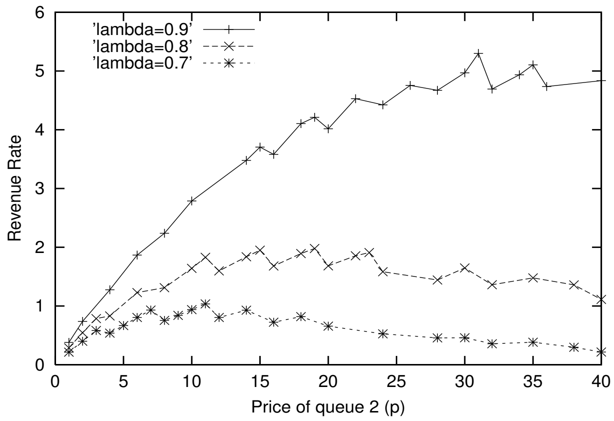

Fig. 7. Revenue rate as a function of p for different λ; a_1=0.8, $a_2 = 0.3$

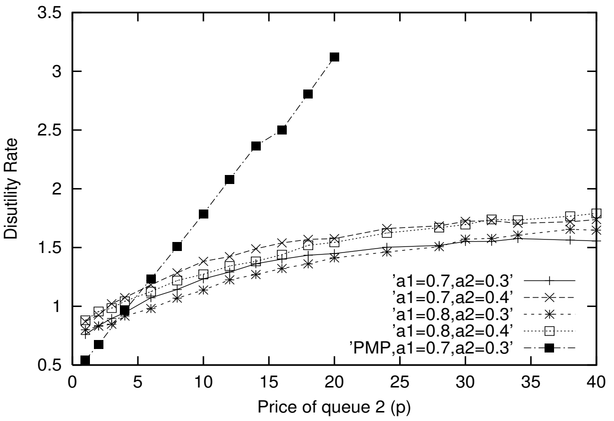

Fig. 6. Disutility rate as a function of p for different a_1 and a_2; λ=0.8.

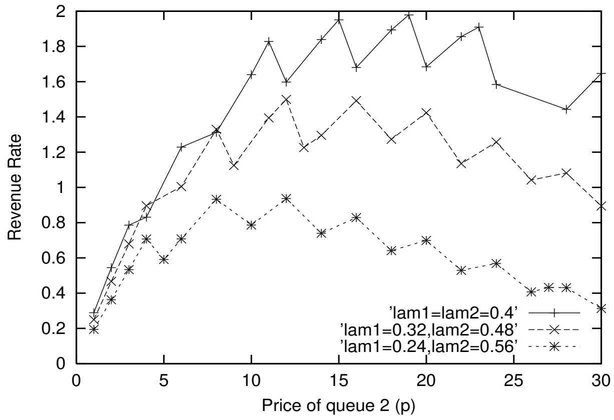

Fig. 8. Revenue rate as a function of price for different λ_1 and λ_2; a_1=0.8, $a_2 = 0.3$

seen to be more sensitive to the arrival rate of the class 1 traffic.

V. Discussion

We have analysed a DiffServ node operating as a join minimum cost queue in the presence of multiclass traffic. Specifically, we have provided a framework that can be used to select appropriate pricing schemes by modeling the tradeoffs between a service provider's gain in terms of its revenue and the customers' cost in terms of a disutility function determined by the price and the delay. Although we have considered a simple form of the disutility function, we believe that the qualitative conclusions that we draw will not vary significantly in the presence of other more sophisticated disutility functions. Specifically, our observation of the significant sensitivity of the revenue to the behavior of the delay sensitive traffic. This means that it is important that the characteristics of that traffic be well modeled.

We are currently investigating the behavior in the presence of more than two classes of traffic and more sophisticated disutility functions. Further work is also aimed at determining the system performance in the presence of balking traffic and finite resources at the queue.

Acknowledgments

We would like to thank Rahul Tandra for discussion on the proofs and for help in computation of the bounds.

References

[1] J. Walrand and P. Varaiya, *High Performance Communication Networks,* 2 Ed, Morgan Kaufman, 1999.
[2] M. Falkner, M. Devetsikiotis, I. Lambadaris, "An Overview of Pricing Concepts for Broadband IP Networks," *IEEE Commun. Surveys*, Sept. 2000, pp. 2-13.
[3] S. Blake, *et al*, "An Architecture for Differential Services," *IETF RFC 2475,* Dec 1998.
[4] A. Odlyzko, "Paris Metro Pricing for the Internet", in the *Proc. of ACM Conf. on Electronic Commerce*, 1999, pp. 140-147.
[5] F. Baumgartner, T. Braun and F. Habegger, "Differentiated Services: A New Approach for Quality of Service in the Internet," in H. van As (ed.): *High Performance Networking,* Kluwer, 1998, pp. 255-274.
[6] R. Gibbens, R. Mason and R. Steinberg, "Internet Service Classes Under Competition," *IEEE JSAC,* vol. 18, no. 12, Dec 2000, pp 2490-2498.
[7] R. Jain, R. Mullen and R. Hausman, "Analysis of Paris Metro Pricing Strategy for QoS with a Single Service Provider," *Proc of IX IWQoS,* Jun 2001.
[8] P. Dube, V. S. Borkar and D. Manjunath, "Differential Join Prices for Parallel Queues: Social Optimality, Dynamic Pricing Algorithms and Application to Internet Pricing," *Proc of IEEE INFOCOM 2002,* Jun 2002.
[9] R. D. Foley and D. R. McDonald, "Join the Shortest Queue: Stability and Exact Asymptotics," Preprint July 1998. `http://www.mathstat.uottawa.ca/profs/mcdonald/homepage/papers/jsq.pdf`.
[10] O. Boxma, G. Koole and Z. Liu, "Queueing Theoretic Solution Methods for Models of Parallel and Distributed Systems," In O. Boxma and G. Koole(eds.).*Performance Evaluation of Parallel and Distributed Systems – Solution Methods,* CWI Tract, 105, pp. 1-24, Amsterdam, 1994.
[11] R. Serfozo, *Introduction to Stochastic Networks,* Springer Verlag, 2000.
[12] G. J. van Houtum, I. J. B. F. Adan, J. Wessesl and W. H. M. Zijm, "Performance Analysis of Parallel Identical Machines with a Generalized Shortest Queue Arrival Mechanism," *OR Spektrum,* vol. 23, 2001, pp. 411-428.
[13] M. F. Neuts, *Matrix Geometric Solutions in Stochastic Models,* Johns Hopkins University Press, Baltimore MD, 1981.