2011 50th IEEE Conference on Decision and Control and
European Control Conference (CDC-ECC)
Orlando, FL, USA, December 12-15, 2011

Load Balancing and Routing Games with Admission Price

Tejas Bodas — IIT Bombay, INDIA
Ayalvadi Ganesh — University of Bristol, UK
D. Manjunath — IIT Bombay, INDIA

Abstract— **We consider load balancing with routing games in a multiclass traffic environment. The servers are M/M/1 type servers and charge an admission price to each customer that joins the queue for service. Service requirements of all arriving customers are i.i.d. and they can receive service from any of the servers. Customers also have a waiting time cost that is proportional to their expected waiting times. Arrivals are from a multiclass population with the different classes differing in the their waiting time costs and having different arrival rates. In this paper we consider the following two load balancing schemes. (1) Both classes are non atomic; each arriving customer independently chooses one of the servers with a probability that optimizes an individual objective function. (2) One of the classes has a dispatcher that routes customers of that class to the servers with probabilities that minimize the total cost for that class; customers of the other class choose a server like in the first scheme. We analyze the equilibrium behavior of both the systems. We also describe a system that can be used to bound the price of anarchy in such systems.**

I. Introduction

We consider a multi server system of non identical servers that uses admission price to achieve load balancing. The system is open—customers arrive into the system and leave after receiving the required service. Traffic is non elastic, i.e., each class of customers has an inherent arrival rate and every arrival receives service. Customers form a multi class population; each class has a unique cost function with expected delay and admission price as variables. Arriving customers do not know the instantaneous queue lengths but know the admission price and system performance expressed as the expected delay at the queue of each server. Arriving customers are independently routed to a server with probabilities that optimize a prescribed cost function.

In this paper we introduce different levels of 'centralization' of the load balancing routing policy and compare their performances under a cost model. Such schemes were first considered in [1], where a system with single class traffic is served by a set of M/G/1 queues but there is no admission price to the queues. A waiting cost proportional to the sojourn time is associated with each queue (and not a class). An arriving customer joins a queue using an individually optimal policy which leads to a Wardrop equilibrium for the system. In [1] it is shown that the socially optimal allocation, is not the same as the allocation at Wardrop equilibrium. In [2], the optimal allocation is compared with that at Wardrop equilibrium and an upper bound on the Price of Anarchy (PoA) is obtained for a system with single class traffic and M/M/1 queues with identical waiting costs. Multi-class traffic was considered in [3] where a class is distinguished by the service time distributions. Each queue is served according to the PS discipline and each customer has a waiting cost proportional to the waiting time in the queue. It is shown that an optimal allocation of customers to the servers is independent of the knowledge of the customer service requirements. The PoA for the individually optimal joining scheme is also calculated.

A more centralized queue-join policy than the individually optimum policies of [1], [2], [3] is in the use of dispatcher for each class. Here, the dispatcher allocates its customers among the queues to minimize the expected delay to its customers. Such a system is considered in [4]. A Class i dispatcher allocates Class i traffic in such a way that the expected waiting time of Class i customers is minimized. It is shown in [4] that a Nash equilibrium (with traffic allocation as the strategy) between the dispatchers exists and the PoA for the system is also obtained. An even more centralized approach would be to use a global objective function and allocate the probabilities for each class-queue pair as in [5]. Here, each queue becomes an M/G/1 queue under a probabilistic allocation and the allocation to minimize the mean waiting time is analyzed in [5].

In the preceding discussion, we see that each customer experiences a waiting cost that is proportional to the sojourn time. The above models assume that this cost depends on the queue and not on the customer class. An alternative pricing mechanism is that of an admission price, e.g., the Paris Metro pricing scheme [6], [7] and Tirupati pricing schemes [8], [9]. Here each queue has a different admission price and an admission price is charged to each customer that joins the queue. A cost function with expected waiting time and admission price is used by each customer to determine the individually optimum queue to join. In such a system, it would be reasonable to have a waiting cost that depends on the customer class and not necessarily on the queue. In this paper we consider such systems.

The rest of the paper is organized as follows. In the next section we consider a non atomic system of two queues with an admission price to each queue. Two classes of traffic arrive into this system. The classes are distinguished by the different waiting time costs. Each arrival randomly chooses one of the queues to minimize an individual objective function. We analyze the equilibrium behavior and characterize the prices that achieve various objectives. In Section III, we consider a similar system except that one

This research was initiated thanks to a Benjamin Meaker Visiting Professorship for D. Manjunath from the Institute of Advanced Study, University of Bristol. Tejas Bodas is supported by an assistantship from the Bharti Centre for Communication at IIT Bombay.

978-1-61284-799-3/11/$26.00 ©2011 IEEE

of the traffic classes has a dispatcher while the other class is a 'non atomic class' in which the arrivals join the queue that minimizes an individual objective function. Once again, we characterize the equilibrium in this system. Finally, in Section IV we consider a genie-based 'ideal model' and compare the optimum allocation obtained that can be used to obtain bounds on the PoA for each model.

We briefly mention the recent interest in elastic source models analyzed in [10]. In these systems an aggregate utility function is associated with each class of traffic and there is an admission price for each arrival. Conditions on the admission prices that would lead to optimal behavior of the system are provided. Such a system for multi-class traffic are also analyzed in, among others, [11]. We will not consider such systems in this paper.

A. Model Overview, Notation and Preliminaries

Customers from two different classes are serviced by queues at two servers. Customers of Class i arrive according to a stationary Poisson process of rate λ_i and Server j is an exponential server of rate μ_j with $i,j = 1,2$; $c_j \geq 0$ is the admission price charged at the queue of Server j; β_i is the cost per unit waiting time incurred by a Class i customer; we assume $\beta_1 > \beta_2$. Also without loss of generality we assume $c_2 = 0$. An arriving Class i customer joins Server 1 with probability p_i independent of all other customers. p_i is determined by the optimization model; let $q_i = 1 - p_i$. Since arriving customers choose the server randomly and independently of the other customers, we have 2 M/M/1 queues in the system. Let γ_j be the total arrival rate to Server j; we have $\gamma_1 = p_1\lambda_1 + p_2\lambda_2$ and $\gamma_2 = q_1\lambda_1 + q_2\lambda_2$, and the expected waiting time in the queue of Server j is $D_j(\gamma_j) := \frac{1}{\mu_j - \gamma_j}$. Thus the expected total cost, sum of the admission price and the expected waiting cost, incurred by a Class i customer receiving service at Server 1 and at Server 2 are, respectively,

$$\delta_{i1} \;=\; c + \beta_i D_1(\gamma_1) \qquad \text{and} \qquad \delta_{i2} \;=\; \beta_i D_2(\gamma_2),$$

the expected total cost for a Class i customer, denoted by δ_i will be $\delta_i = p_i\delta_{i1} + q_i\delta_{i2}$ and the expected total cost per customer will be $\delta_s = (\lambda_1\delta_1 + \lambda_2\delta_2)\,/\,(\lambda_1 + \lambda_2)$.

II. Two Non Atomic Classes

Recall that in this model, each customer performs an individual optimization and joins the queue that minimizes its expected cost. This decision is based on the admission price and the mean waiting time at the servers. The customers do not know the instantaneous queue lengths. Thus, at equilibrium, if Class i traffic is using two different servers at equilibrium, then the expected total cost at the two servers should be the same. If this is not the case, then some Class i customers would move to a cheaper server indicating the traffic earlier was not at equilibrium. At equilibrium, the servers that are not used by a class of customers have a higher expected cost than that of servers that have a non zero arrival rate from that class. The equilibrium is clearly a Wardrop equilibrium.

A. Traffic Distribution at Wardrop Equilibrium

We begin with the following theorem.

***Theorem** 1:* Only one of the following is true

1) $p_1 = 1$ and $1 \;\geq\; p_2 \;\geq\; 0$.
2) $1 \;>\; p_1 \geq 0$ and $p_2 = 0$.

Proof: Under Wardrop equilibrium we have the following. For Class i, $i = 1,2$, the following is true.

$$\begin{array}{lll} p_i = 0 & \text{iff} & c + \beta_i D_1(\gamma_1) \;>\; \beta_1 D_2(\gamma_2). \\ 0 \;<\; p_i \;<\; 1 & \text{iff} & c + \beta_i D_1(\gamma_1) \;=\; \beta_i D_2(\gamma_2). \\ p_i = 1 & \text{iff} & c + \beta_i D_1(\gamma_1) \;<\; \beta_i D_2(\gamma_2). \end{array} \tag{1}$$

From the preceding inequalities,

- $p_1 = 0$ iff $c \;>\; \beta_1\left(D_2(\gamma_2) - D_1(\gamma_1)\right)$. Since $\beta_1 > \beta_2$ by assumption, $c \;>\; \beta_2\;\left(D_2(\gamma_2) - D_1(\gamma_1)\right)$ which implies $p_2 = 0$.
- $p_2 = 1$ iff $c \;<\; \beta_2(D_2(\gamma_2) - D_1(\gamma_1)) \;<\; \beta_1(D_2(\gamma_2) - D_1(\gamma_1))$. This means that $p_2 = 1$ implies $p_1 = 1$.
- $0 \;<\; p_2 < 1$ iff $c = \beta_2(D_2(\gamma_2) - D_1(\gamma_1))$. Because $\beta_1 > \beta_2$, this means $c \neq \beta_1(D_2(\gamma_2) - D_1(\gamma_1))$. This in turn means that with $0 \;<\; p_2 < 1$, we cannot have $0 \;<\; p_1 \;<\; 1$. However, $c \;=\; \beta_2(D_2(\gamma_2) - D_1(\gamma_1))$ implies $c < \beta_1(D_2(\gamma_2) - D_1(\gamma_1))$. Therefore, if $0 \;<\; p_2 < 1$, then $p_1 = 1$.
- $0 < p_1 < 1$ iff $c + \beta_1 D_1(\gamma_1) \;=\; \beta_1 D_2(\gamma_2)$. This also implies $c > \beta_2(D_2(\gamma_2) - D_1(\gamma_1))$. In this case only $p_2 = 0$ is possible. □

Theorem 1 determines the pattern of traffic flows in the two queues at Wardrop equilibrium for different values c, λ_i and μ_i. The following five regimes can now be defined—(1) Regime 1 for which $p_1 = p_2 = 1$, (2) Regime 2 for which $p_1 = 1$ and $0 < p_2 < 1$, (3) Regime 3 for which $p_1 = 1$ and $p_2 = 0$, (4) Regime 4 for which $0 < p_1 < 1$ and $p_2 = 0$, and (5) Regime 5 for which $p_1 = 0$ and $p_2 = 0$.

We first analyze the effect of c on the traffic distribution by fixing the arrival rates λ_i and the service rates μ_i to obtain the change in the equilibrium traffic when the admission price is increased from 0. Of course, a regime is feasible only if both the queues are stable, i.e., $\mu_i \;>\; \gamma_i$ for $i \;=\; 1,2$. This characterization is then used to analyze c as a control parameter that (1) maximizes the revenue rate, and (2) minimizes the waiting time costs.

In Regime 1, $p_1 = p_2 = 1$ and hence $\gamma_1 = \lambda_1 + \lambda_2$ and $\gamma_2 = 0$. Here, we will need $c < \beta_2(D_2(0) - D_1(\lambda_1 + \lambda_2))$. Let $c_1 := \beta_2\left(D_2(0) - D_1(\lambda_1 + \lambda_2)\right)$, i.e.,

$$c_1 := \frac{\beta_2}{\mu_2} - \frac{\beta_2}{\mu_1 - \lambda_1 - \lambda_2} \tag{2}$$

Thus, if $\mu_1 > \lambda_1 + \lambda_2$ and $0 \;<\; c \;<\; c_1$, then the Wardrop equilibrium will be in Regime 1.

Regime 3 requires $p_1 = 1$ and $p_2 = 0$, i.e., $\gamma_1 = \lambda_1$ and $\gamma_2 = \lambda_2$. This leads to the following condition on c.

$$\beta_2(D_2(\lambda_2) - D_1(\lambda_1)) \;<\; c \;<\; \beta_1(D_2(\lambda_2) - D_1(\lambda_1))$$

Let

$$c_2 := D_2(\lambda_2) - D_1(\lambda_1) = \frac{1}{\mu_2 - \lambda_2} - \frac{1}{\mu_1 - \lambda_1} \tag{3}$$

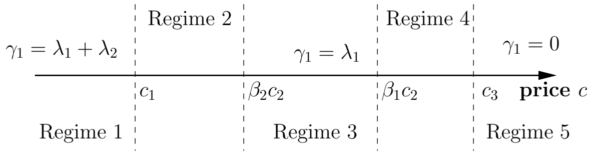

Fig. 1. The operating regimes as c is increased from 0. In addition to the requirement on c, the stability conditions also need to be satisfied. c_1, c_2 and c_3 are as in Eqns. 2, 3 and 4 respectively.

Thus if $\mu_1 > \lambda_1$, $\mu_2 > \lambda_2$ and $0 < \beta_2 c_2 < c < \beta_1 c_2$, then the Wardrop equilibrium will operate in Regime 3. The condition $c_2 > 0$ is possible only if $\mu_1 - \lambda_1 > \mu_2 - \lambda_2$ which thus becomes a necessary condition for this regime to be possible.

If the equilibrium is in Regime 2, then $p_1 = 1$ and $0 < p_2 < 1$ and $\gamma_1 = \lambda_1 + p_2\lambda_2$ and $\gamma_2 = q_2\lambda_2$. For Regime 2 to to be possible for all $p_2 \in [0, 1]$, we need $\mu_1 > \lambda_1 + \lambda_2$ and $\mu_2 > \lambda_2$. Since $p_1 = 1$ and $0 < p_2 < 1$, for the queue to be in this regime we require

$$c \;=\; \frac{\beta_2}{\mu_2 - q_2\lambda_2} - \frac{\beta_2}{\mu_1 - \lambda_1 - p_2\lambda_2.}$$

Observe that as we move from Regime 1 ($p_2 = 1$) to Regime 3 ($p_2 = 0$) through Regime 2, p_2 decreases from 1 to 0. This happens as c is increased from c_1 to $\beta_2 c_2$. We can show that $c_1 \le \beta_2 c_2$ with equality if $\lambda_2 = 0$.

For equilibrium in Regime 5, we need $p_1 = p_2 = 0$, i.e., $\gamma_1 = 0$ and $\gamma_2 = \lambda_1 + \lambda_2$. This requires $c > \beta_1((D_2(\lambda_1 + \lambda_2) - D_1(0)))$. Define $c_3 := \beta_1(D_2(\lambda_1 + \lambda_2) - D_1(0))$, i.e.,

$$c_3 := \frac{\beta_1}{\mu_2 - \lambda_1 - \lambda_2} - \frac{\beta_1}{\mu_1}. \tag{4}$$

Thus the equilibrium will be in Regime 5 if $c_3 > 0$, $\mu_2 > \lambda_1 + \lambda_2$ (stability condition), and $c > c_3$.

Operating in Regime 4 requires $p_2 = 0$ and $0 < p_1 < 1$ which in turn means $\gamma_1 = p_1\lambda_1$ and $\gamma_2 = q_1\lambda_1 + \lambda_2$. Thus, a particular (p_1, p_2) satisfying $0 < p_1 < 1$ and $p_2 = 0$ is achieved with

$$c \;=\; \beta_1\left(\frac{1}{\mu_2 - q_1\lambda_1 - \lambda_2} - \frac{1}{\mu_1 - p_1\lambda_1}\right).$$

Thus, as c increases from $\beta_1 c_2$ to c_3, p_1 decreases from 1 to 0. The equilibrium will be in Regime 4 if $0 < \beta_1 c_2 < c < c_3$ and the stability condition $\mu_1 > \lambda_1$ and $\mu_2 > \lambda_1 + \lambda_2$.

Figure 1 summarizes the preceding discussion.

B. Maximizing Revenue Rate

We now characterize the c^* that maximizes the revenue rate for the two-queue system. Observe that as c is increased from $c_1 > 0$ to $\beta_2 c_2$, γ_1 decreases from $\lambda_1 + \lambda_2$ to λ_1, whereas as c is increased form $\beta_1 c_2$ to c_3, γ_1 decreases from λ_1 to 0. Clearly, if Regime 1 is the preferred regime for equilibrium, then $c = c_1$ maximizes the revenue rate. Also, $c > c_3$ results in $\gamma_1 = 0$ and would provide no revenue. Thus for $c_1 \le c \le c_3$, and $0 < \gamma_1 < \lambda_1 + \lambda_2$, there is a non zero revenue rate.

Let $c_k(\gamma_1)$ be the admission price that achieves an arrival rate of γ_1 to Queue 1 in Regime k, c_k^* be the c that maximizes the revenue when operating in Regime k and $\gamma_{1,k}^*$ the arrival rate into Queue 1 when $c = c_k^*$. Clearly, $c_1^* = c_1$ and $c_3^* = \beta_1 c_2$ and $\gamma_{1,1}^* = (\lambda_1 + \lambda_2)$ and $\gamma_{1,3}^* = \lambda_1$. Further,

$$\begin{aligned} c_2(\gamma_1) &= \beta_2\left(\frac{1}{\mu_2 - \gamma_2} - \frac{1}{\mu_1 - \gamma_1}\right) \\ c_4(\gamma_1) &= \beta_1\left(\frac{1}{\mu_2 - \gamma_2} - \frac{1}{\mu_1 - \gamma_1}\right) \end{aligned} \tag{5}$$

It can be shown that $\gamma_1 c_2(\gamma_1)$ is concave in γ_1 if $\mu_i > \gamma_i$ for $i = 1, 2$; hence $c_1 \le c_2^* \le \beta_2 c_2$. Let $\hat{\gamma}_{1,2}$ be a feasible solution of

$$\frac{\mu_2 - \lambda_1 - \lambda_2}{(\mu_2 - \lambda_1 - \lambda_2 + \hat{\gamma}_{1,2})^2} = \frac{\mu_1}{(\mu_1 - \hat{\gamma}_{1,2})^2}$$

which is the maximizing condition obtained by after differentiating $\gamma_1 c_2(\gamma_1)$ w.r.t γ_1 and equating to zero. Solving the quadratic equation for $\hat{\gamma}_{1,2}$, we require $(\mu_2 - \lambda_1 - \lambda_2) > 0$ for the roots to be real; this will render one of the roots infeasible because it will require $\hat{\gamma}_{1,2} > \mu_1$. Thus the feasible solution will be $\hat{\gamma}_{1,2} = \frac{\mu_1 - \sqrt{\mu_1(\mu_2 - \lambda_1 - \lambda_2)}}{1 + \sqrt{\frac{\mu_1}{(\mu_2 - \lambda_1 - \lambda_2)}}}$. $\hat{\gamma}_{1,2}$ is feasible if $c_1 < c_2(\hat{\gamma}_{1,2}) < \beta_2 c_2$. Using the concavity of $\gamma_1 c_2(\gamma_1)$, we obtain

$$c_2^* = \begin{cases} c_1 & \text{if } c_1(\lambda_1 + \lambda_2) > \lambda_1\beta_2 c_2 \;\&\; c_2(\hat{\gamma}_{1,2}) \notin (c_1, \beta_2 c_2) \\ \beta_2 c_2 & \text{if } c_1(\lambda_1 + \lambda_2) < \lambda_2\beta_2 c_2 \;\&\; c_2(\hat{\gamma}_{1,2}) \notin (c_1, \beta_2 c_2) \\ c_2(\hat{\gamma}_{1,2}) & \text{otherwise} \end{cases}$$

c_4^* can be obtained analogously and c^* is obtained as $\arg\max_{c_k : 1 \le k \le 4}\{\gamma_{1,k}^* c_k^*\}$.

C. Minimizing the Waiting Time Cost in the System

Let us now obtain the $c^\star$, equivalently the $\gamma_1^\star$, that minimizes the waiting time cost at equilibrium. Let Δ_i be the waiting time cost for a Class i customer; $\Delta_i = p_i\beta_i/(\mu_1 - \gamma_1) + (1 - p_i)\beta_i/(\mu_2 - \gamma_2)$. The system waiting time cost when the total arrival rate into Queue 1 is γ_1 is denoted by $\Delta_{\gamma_1} = (\Delta_1\lambda_1 + \Delta_2\lambda_2) \,/\, (\lambda_1 + \lambda_2)$. As before, let $c_k^\star$ be the c that minimizes Δ_{γ_1} when operating in Regime k and $\gamma_{1,k}^\star$ the arrival rate into Queue 1 when $c = c_k^\star$. In Regimes 1, 3 and 5, the traffic distribution does not vary with c and $\gamma_{1,1}^\star = \lambda_1 + \lambda_2$, $\gamma_{1,3}^\star = \lambda_1$, and $\gamma_{1,5}^\star = 0$. We can thus use $c_1^\star = 0$, $c_3^\star = \beta_2 c_2$, and $c_5^\star = c_3$. In Regime 2 $p_1 = 1$, and hence

$$\Delta_{\gamma_1} = \frac{\beta_1\lambda_1 + p_2\beta_2\lambda_2}{(\mu_1 - \gamma_1)(\lambda_1 + \lambda_2)} + \frac{(\beta_1\lambda_1 + \beta_2\lambda_2) - (\beta_1\lambda_1 + p_2\beta_2\lambda_2)}{(\mu_2 - \gamma_2)(\lambda_1 + \lambda_2)}$$

Let $\hat{\gamma}_{1,2}$ be a feasible solution of

$$\frac{\beta_2\lambda_2\mu_2}{(\mu_2 - \gamma_2)^2} = \frac{\beta_2\lambda_2(\mu_1 - \lambda_1) + \beta_1\lambda_1\lambda_2}{(\mu_1 - \gamma_1)^2}.$$

which is the minimizing condition obtained by differentiating Δ_{γ_1} w.r.t p_2 with $p_1 = 1$, and equating to zero. Solving for $\hat{\gamma}_{1,2}$, and using $z = (\beta_2\lambda_2(\mu_1 - \lambda_1) + \beta_1\lambda_1\lambda_2) \,/\, (\beta_2\lambda_2\mu_2)$, we have two feasible roots given by $\hat{\gamma}_{1a,2} = (\mu_1 - (\mu_2 - \lambda_1 - \lambda_2)\sqrt{z}) \,/\, (1 + \sqrt{z})$ and $\hat{\gamma}_{1b,2} =$

$(\mu_1 + (\mu_2 - \lambda_1 - \lambda_2)\sqrt{z}) \,/\, (1 - \sqrt{z})$. The two roots are feasible if $0 \leq \hat{\gamma}_{1a,2}, \hat{\gamma}_{1b,2} \leq \lambda_1 + \lambda_2$.

It can be shown that Δ_{γ_1} is convex in p_2; hence

$$\gamma^\star_{1,2} = \begin{cases} \hat{\gamma}_{1a,2} & \text{if } \lambda_1 \leq \hat{\gamma}_{1a,2} \leq \lambda_1 + \lambda_2 \text{ \& } \Delta_{\hat{\gamma}_{1a,2}} < \Delta_{\hat{\gamma}_{1b,2}} \\ \hat{\gamma}_{1b,2} & \text{if } \lambda_1 \leq \hat{\gamma}_{1b,2} \leq \lambda_1 + \lambda_2 \text{ \& } \Delta_{\hat{\gamma}_{1b,2}} < \Delta_{\hat{\gamma}_{1a,2}} \\ \lambda_1 & \text{if } \Delta_{\lambda_1} < \Delta_{\lambda_1+\lambda_2} \text{ \& } \hat{\gamma}_{1a,2}, \hat{\gamma}_{1b,2} \notin [\lambda_1, \lambda_1 + \lambda_2] \\ \lambda_1 + \lambda_2 & \text{otherwise} \end{cases}$$

The corresponding $c^\star_2$ can be obtained from (5).

$\gamma^\star_{1,4}$ is obtained analogously as

$$\gamma^\star_{1,4} = \begin{cases} \hat{\gamma}_{1c,2} & \text{if } 0 \leq \hat{\gamma}_{1c,2} \leq \lambda_1 \text{ and } \Delta_{\hat{\gamma}_{1c,2}} < \Delta_{\hat{\gamma}_{1d,2}} \\ \hat{\gamma}_{1d,2} & \text{if } 0 \leq \hat{\gamma}_{1d,2} \leq \lambda_1 \text{ and } \Delta_{\hat{\gamma}_{1d,2}} < \Delta_{\hat{\gamma}_{1c,2}} \\ \lambda_1 & \text{if } \Delta_{\lambda_1} < \Delta_0 \text{ and } \hat{\gamma}_{1c,2}, \hat{\gamma}_{1d,2} \notin [\lambda_1, \lambda_1 + \lambda_2] \\ 0 & \text{otherwise.} \end{cases}$$

Here, $\hat{\gamma}_{1c,2} = (\mu_1 - (\mu_2 - \lambda_1 - \lambda_2)\sqrt{w}) \,/\, (1 + \sqrt{w})$, $\hat{\gamma}_{1d,2} = (\mu_1 + (\mu_2 - \lambda_1 - \lambda_2)\sqrt{w}) \,/\, (1 - \sqrt{w})$, and $w = (\beta_1 \lambda_1 \mu_1) \,/\, (\beta_2 \lambda_1 \lambda_2 + \beta_1 \lambda_1 (\mu_2 - \lambda_2))$.

Finally, $\gamma^\star_1$ is obtained as

$$\arg \min_{\gamma^\star_{1,k} : 1 \leq k \leq 5} \{\Delta_{\gamma^\star_{1,k}}\}.$$

III. A Dispatcher and a Non Atomic Class

We now consider the system where Class 1 customers are routed by a dispatcher that chooses the routing fraction to optimize the total cost for Class 1 customers. Class 2 customers are routed to optimize and individual objective function. In our analyzes below we will often fix p_1 and let Class 2 customers be routed to achieve *Class 2 equilibrium.* This equilibrium is clearly distinct from the *system equilibrium* at which p_1 and p_2 are in mutual equilibrium, i.e., p_1 is the optimum for the corresponding p_2 and p_2 achieves the Class 2 equilibrium for the corresponding p_1.

If p_1 is fixed and Class 2 equilibrium is achieved with p_2, then the total cost for a Class 1 customer is

$$\delta_1(p_1) = p_1 \left(c + \frac{\beta_1}{\mu_1 - p_1\lambda_1 - p_2\lambda_2} \right) + \left(\frac{\beta_1(1 - p_1)}{\mu_2 - q_1\lambda_1 - q_2\lambda_2} \right). \tag{6}$$

It can be shown that when the queues are stable, $\delta_1(p_1)$ is convex in p_1.

Like in the previous section we begin by analyzing the system equilibrium as c is increased from 0. Also, the definitions of c_1, and c_2 and c_3 are also as in the previous section. We look on the same scale of the admission price as that of the non-atomic model, i.e., we start with $c > 0$ and increase c beyond $c_1, \beta_1 c_2$ and c_3 and observe the resulting system equilibrium between the Class 1 and Class 2 traffic where the Class 1 traffic flow is now regulated by the dispatcher.

***Property** 1:* At system equilibrium, if $c < c_1$, then $p_2 = 1$ and if $c > \frac{\beta_2}{\beta_1} c_3$, then $p_2 = 0$.

Proof:

$$\begin{aligned} c \quad < \quad c_1 \;&=\; \frac{\beta_2}{\mu_2} - \frac{\beta_2}{\mu_1 - \lambda_1 - \lambda_2} \\ &\leq\; \left(\frac{\beta_2}{\mu_2 - q_1\lambda_1} - \frac{\beta_2}{\mu_1 - p_1\lambda_1 - \lambda_2} \right) \end{aligned}$$

which in turn implies $p_2 = 1$. Similarly, for $c > \frac{\beta_2}{\beta_1} c_3$,

$$\begin{aligned} c \quad > \tfrac{\beta_2}{\beta_1} c_3 \quad &= \quad \frac{\beta_2}{\beta_1} \left(\frac{\beta_1}{\mu_2 - \lambda_1 - \lambda_2} - \frac{\beta_1}{\mu_1} \right) \\ &\geq \quad \frac{\beta_2}{\mu_2 - q_1\lambda_1 - \lambda_2} - \frac{\beta_2}{\mu_1 - p_1\lambda_1}. \end{aligned}$$

which implies $p_2 = 0$ for Class 2 at system equilibrium. □

***Property** 2:* For $0 < c < c_1$, at system equilibrium the following are true.

1) $p_1 \neq 0$.
2) If $c \leq \dfrac{\beta_1}{\mu_2} - \dfrac{\beta_1(\mu_1 - \lambda_2)}{(\mu_1 - \lambda_1 - \lambda_2)^2} < c_1$,then $p_1 = 1$.
3) If $\dfrac{\beta_1}{\mu_2} - \dfrac{\beta_1(\mu_1 - \lambda_2)}{(\mu_1 - \lambda_1 - \lambda_2)^2} < c < c_1$, then $0 \leq p_1 \leq 1$.

Proof: From Property 1, for $c < c_1$, $p_2 = 1$. We will show that with $p_2 = 1$, the cost for Class 1 is lower with $p_1 = 1$ than with $p_1 = 0$. Thus $p_1 = 0$ will not be a feasible system equilibrium. For $p_1 = 1$, $\delta_1 = c + \dfrac{\beta_1}{\mu_1 - \lambda_1 - \lambda_2}$ and for $p_1 = 0$, $\delta_1 = \dfrac{\beta_1}{\mu_2 - \lambda_1}$. Using the assumption that $\beta_2 < \beta_1$, and the definition of c_1, for $c < c_1$, we can write,

$$\begin{aligned} c + \frac{\beta_1}{\mu_1 - \lambda_1 - \lambda_2} \quad &< \quad c_1 + \frac{\beta_1}{\mu_1 - \lambda_1 - \lambda_2} \\ &< \quad \frac{\beta_1}{\mu_2} \quad < \quad \frac{\beta_1}{\mu_2 - \lambda_1}. \end{aligned}$$

This proves part 1. Let $\delta_1'(p_1)$ be the derivative of δ_1 (in (6)) w.r.t p_1. From the convexity of δ_1, and using the fact that at system equilibrium, $p_1 \neq 0$,, $\delta_1'(1) > 0$ implies that at system equilibrium $0 < p_1 < 1$. Conversely, $\delta_1'(1) \leq 0$ implies $p_1 = 1$. We can show that, if $c > \dfrac{\beta_1}{\mu_2} - \dfrac{\beta_1(\mu_1 - \lambda_2)}{(\mu_1 - \lambda_1 - \lambda_2)^2}$, $\delta_1'(1) > 0$. And $c \leq \dfrac{\beta_1}{\mu_2} - \dfrac{\beta_1(\mu_1 - \lambda_2)}{(\mu_1 - \lambda_1 - \lambda_2)^2}$, implies $\delta'(1) \leq 0$. Parts 2 and 3 are thus proved. □

We now explain the system equilibrium for $c > \frac{\beta_2}{\beta_1} c_3$.

***Property** 3:* For $0 < \frac{\beta_2}{\beta_1} c_3 < c$, the system equilibrium p_1 satisfies $0 \leq p_1 \leq 1$.

Proof: Like in the proof of the preceding property, consider $\delta_1'(p_1)$, the derivative of the RHS of (6) w.r.t p_1 but with $p_2 = 0$. Arguing as before, $\delta_1'(0) > 0$ implies $p_1 = 0$. We can show that this if $c > \dfrac{\beta_1(\mu_2 - \lambda_2)}{(\mu_2 - \lambda_1 - \lambda_2)^2} - \dfrac{\beta_1}{\mu_1}$.

Similarly, $\delta_1'(1) < 0$ implies $p_1 = 1$ which is true when $c < \dfrac{\beta_1}{(\mu_2 - \lambda_2)} - \dfrac{\beta_1\mu_1}{(\mu_1 - \lambda_1)^2}$.

Finally, for $0 < \frac{\beta_2}{\beta_1} c_3 < c$ and

$$\frac{\beta_1}{(\mu_2 - \lambda_2)} - \frac{\beta_1\mu_1}{(\mu_1 - \lambda_1)^2} \quad \leq \quad c \quad \leq \quad \frac{\beta_1(\mu_2 - \lambda_2)}{(\mu_2 - \lambda_1 - \lambda_2)^2} - \frac{\beta_1}{\mu_1}$$

we have $0 < p_1 < 1$ at system equilibrium. □

We now consider the equilibrium traffic distribution when $0 < c_1 < c < \frac{\beta_2}{\beta_1} c_3$. We begin with the following result.

***Lemma** 1:* For $c_1 < c < \frac{\beta_2}{\beta_1} c_3$ and for a fixed, (not necessarily the optimum or equilibrium) p_1, if the equilibrium p_2 of Class 2, satisfies $0 < p_2 < 1$ then the p_2 is unique.

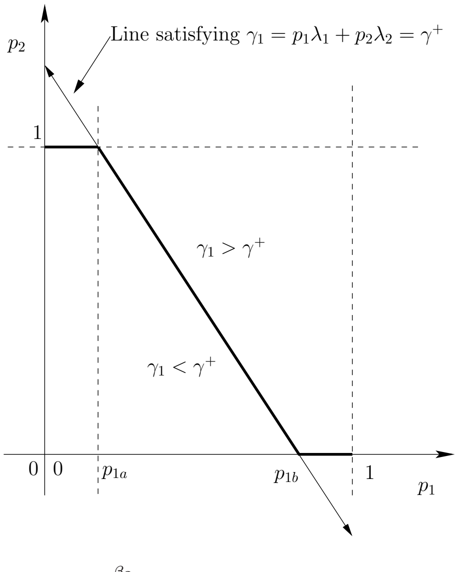

Fig. 2. For $c_1 < c < \frac{\beta_2}{\beta_1}c_3$, and a fixed p_1, the p_2 that achieves Class 2 equilibrium is plotted as a function of p_1.

Proof: Recall that under the conditions of the lemma, at equilibrium, the following equality holds.

$$c + \frac{\beta_2}{\mu_1 - p_1\lambda_1 - p_2\lambda_2} = \frac{\beta_2}{\mu_2 - q_1\lambda_1 - q_2\lambda_2} \qquad (7)$$

The LHS of (7) is monotone increasing and RHS is monotone decreasing in p_1. Further, at $p_1 = 0$ the RHS is larger than LHS while at $p_1 = p_2 = 1$, the LHS is larger than RHS for $c_1 < c < \frac{\beta_2}{\beta_1}c_3$. □

Denote the unique γ_1 obtained in Lemma 1 by γ^+. We now use Lemma 1 to characterize p_2 at Class 2 equilibrium for a fixed p_1, $0 \leq p_1 \leq 1$. For the conditions of Lemma 1, defining $p_{1a} := \frac{\gamma^+ - \lambda_2}{\lambda_1}$, we see that at Class 2 equilibrium, if $p_1 = p_{1a}$ and $p_2 = 1$, then we will have $\gamma_1 = \gamma^+$. This means that if $p_1 < p_{1a}$ then we have $\gamma_1 < \gamma^+$ and hence

$$c + \frac{\beta_2}{\mu_1 - \gamma_1} < \frac{\beta_2}{\mu_2 - \gamma_2}.$$

From the Class 2 equilibrium conditions this implies that $p_2 = 1$ when p_1 is fixed to $p_1 < p_{1a}$. Defining $p_{1b} = \gamma^+/\lambda_1$, and arguing similarly, we see that if p_1 is fixed at $p_1 > p_{1b}$, then at Class 2 equilibrium $\gamma_1 > \gamma^+$ and

$$c + \frac{\beta_2}{\mu_1 - \gamma_1} > \frac{\beta_2}{\mu_2 - \gamma_2},$$

and the Class 2 equilibrium conditions implies that $p_2 = 0$.

Summarizing the preceding discussion, for a fixed p_1, at Class 2 equilibrium, if $0 \leq p_1 \leq p_{1a} \leq 1$, then $p_2 = 1$ and if $0 \leq p_{1b} \leq p_1 \leq 1$, then $p_2 = 0$. Also, for $p_{1a} < p_1 < p_{1b}$ with $\gamma_1 = \gamma^+$ we have $0 < p_2 < 1$. This is summarized in Fig. 2.

We remind the reader that the preceding discussion applies to each c satisfying $c_1 < c < \frac{\beta_2}{\beta_1}c_3$. We now observe that if $p_{1a} \leq p_1 \leq p_{1b}$, then we will have $\gamma_1 = \gamma^+$ and hence $c + \frac{\beta_2}{\mu_1 - \gamma^+} = \frac{\beta_2}{\mu_2 - \lambda_1 - \lambda_2 + \gamma^+}$. This also implies that $c + \frac{\beta_1}{\mu_1 - \gamma^+} < \frac{\beta_1}{\mu_2 - \lambda_1 - \lambda_2 + \gamma^+}$. Now as $\delta_1 = p_1\left(c + \frac{\beta_1}{\mu_1 - \gamma^+}\right) + \left(\frac{\beta_1(1-p_1)}{\mu_2 - \lambda_1 - \lambda_2 + \gamma^+}\right)$ it is obvious that for $p_{1a} \leq p_1 \leq p_{1b}$, δ_1 is minimized at $p_1 = p_{1b}$ with the corresponding $p_2 = 0$. We thus have the following property.

Property 4: For $0 < c_1 < c < \frac{\beta_2}{\beta_1}c_3$ the equilibrium (p_1, p_2) satisfies one of the following conditions.

1) $0 \leq p_1 \leq p_{1a} \leq 1$ and $p_2 = 1$.
2) $0 \leq p_{1b} \leq p_1 \leq 1$ and $p_2 = 0$.
3) If $p_{1b} \notin [0, 1]$ then $p_1 = 1$ and $p_2 = \frac{\gamma^+ - \lambda_1}{\lambda_2}$.

The proof follows from Fig. 2 and the preceding discussion. Thus, corresponding to every c in the said range, of the three conditions above, the one which has the least δ_1 will prevail as the system equilibrium.

IV. Towards a Bound on the Price of Anarchy

The price of anarchy (PoA) is the ratio of the value of the objective function at the social optimum routing to that at *equilibrium.* The social optimum routing policy of the customers is not known. Although the Bernoulli routing of [5] is a centralized scheme and optimizes the system cost, it is not known if this is indeed the socially optimal routing. In this section we describe a two-class, two-server system in which the total expected cost per customer will be lower than that under a socially optimal scheme and hence can be used to obtain an upper bound on the PoA for the various load balancing schemes that we have outlined in this paper.

Consider the following system. Arriving customers of Class i wait in queue Q_i before receiving service. Both the servers complete 'virtual services' irrespective of whether the corresponding queue is non-empty. Thus, the virtual service process at server j is a Poisson(μ_j) process. (By the memoryless property of the exponential distribution, it makes no difference whether we start serving before a customer arrives at the queue, or wait till it arrives.) Customer arrivals form a Poisson process of rate λ_i for class i. After completion of each service by a server, if there are waiting customers in any of the queues then, the completed service is 'allocated' to one of them according to a policy that we describe below. The customer then leaves the system.

We now motivate the allocation policy. Consider a service completion epoch t. At this epoch, let M_1 and M_2 be the number of Class 1 and Class 2 customers in queues Q_1 and Q_2 respectively. If we choose to allocate the service to a waiting customer, then it has to pay the corresponding admission price. On the other hand, if this service is discarded, then the an opportunity cost resulting from an increase in the waiting time costs is incurred. If the service is not allocated the opportunity to send out a waiting customer will arrive only at the next service completion epoch. And the customers that are already in the queue will experience at least that much extra waiting time. Further, those customers that arrive before the next service completion epoch will have to wait more than they would have if this service is allocated to a waiting customer. We can now derive our service allocation policy—a service completion is allocated to a waiting customer if the additional waiting time cost is more than the admission price for that queue. Also the allocation will be to the customer that will cause a maximum reduction in the system cost. We now

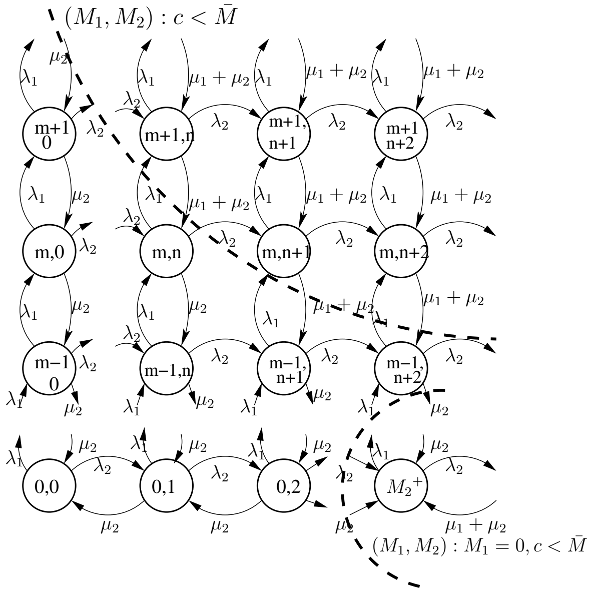

Fig. 3. The Markov chain representation for the bounding system.

calculate the waiting time cost when a service completion is not allocated.

The time until the next service completion epoch is exponential with rate $\mu_1 + \mu_2$. Thus the total waiting cost from all customers that are in queue at a service completion epoch is $(\beta_1 M_1 + \beta_2 M_2)/(\mu_1 + \mu_2)$. The extra waiting time for every customer that arrives between service completion epochs is $1/(\mu_1 + \mu_2)$ and the total cost from such customers will be $(\beta_1 \lambda_1 + \beta_2 \lambda_2)/\left((\mu_1 + \mu_2)^2\right)$. Thus the total cost of missing an opportunity to use a service is

$$\bar{M} = \frac{\beta_1 M_1 + \beta_2 M_2}{\mu_1 + \mu_2} + \frac{\beta_1 \lambda_1 + \beta_2 \lambda_2}{(\mu_1 + \mu_2)^2}$$

Thus we can form the following allocation rule at the completion of a service as follows.

- When Server 2 finishes service,
  - if $M_1 > 0$, then the service is allotted to a Class 1 customer,
  - else if $M_1 = 0$, $M_2 > 0$, then the service is allotted to a Class 2 customer,
  - else the service is discarded.
- When Server 1 finishes service,
  - If $M_1 > 0$ and $c < \bar{M}$, then the service should be allocated to a customer of Class 1,
  - else if $M_1 = 0$ and $M_2 > 0$ and $c < \bar{M}$ then the service is allotted to a Class 2 customer,
  - else the service is discarded.

Since the service allocation is being performed after a service, it can be seen that the system cost (sum of the admission waiting time costs) in this system is less than that in any system with a non anticipatory policy.

The two-dimensional process $(M_1(t), M_2(t))$ is a continuous time Markov chain. Let $\pi_{i,j}$ be the stationary distribution of the Markov chain. Let $\bar{N}_i$ be the expected number of Class i customers. The waiting time cost per unit time of a Class i customer is thus $\bar{N}_i \beta_i$. Let $\mathcal{M}$ be the set of states for which $c < \bar{M}$. Then the rate at which the admission cost is accrued is $c \sum_{ij \in \mathcal{M}} \pi_{ij} \mu_1$. Thus the system cost per unit time is obtained as the sum of the two costs.

While a closed form stationary distribution appears to be rather hard to obtain, a numerical evaluation is immediately possible with a suitable truncation of the state space. A more detailed analysis like in [12], [13] also seem possible.

V. Discussion

We considered multiclass traffic being serviced by non homogeneous servers. We first considered a decentralized model of each customer routing itself to optimize individual objective functions. We then allowed one class to have a dispatcher that routes to optimize for the class. Note that in the second model, our assumption of $\beta_1 > \beta_2$ is not general and an identical analysis can be carried out for $\beta_1 < \beta_2$.

Two other models are also possible. (1) Provide a separate dispatcher for each traffic class. Here the dispatchers compete with each other with their respective class traffic allocation as their strategy. This is similar to the one in [4], except that we now have an admission price into the queues. The analysis will mirror that in [4]. (2) A centralized scheme of a single dispatcher that determines the routing fractions to optimize a global objective. This has similarities to that in [5] except that we assume identical service requirements for each class and there is an admission price.

References

[1] C. H. Bell and S. Stidham, “Individual versus social optimization in the allocation of customers to alternative servers,” *Management Science*, vol. 29, pp. 831–839, 1983.

[2] M. Haviv and T. Roughgarden, “The price of anarchy in an exponential muli-server.,” *Operation Research Letters*, vol. 35, pp. 421–426, 2007.

[3] E. Altman, U. Ayesta, and B. Prabhu, “Load balancing in processor sharing systems,” in *Proc. of 3rd Intnl. Conf. on Perf. Eval. Methodologies and Tools (ValueTools)*, 2008.

[4] U. Ayesta, O. Brun, and B. J. Prabhu, “Price of anarchy in non-cooperative load balancing,” in *Proc. of the IEEE INFOCOM*, 2010, pp. 436–440.

[5] S. C. Borst, “Optimal probabilistic allocation of customer types to servers,” in *Proc. of ACM SIGMETRICS*, Sept. 1995, pp. 116–125.

[6] A. Odlyzko, “Paris Metro pricing for the internet,” in *Proc. of the 1st ACM conference on Electronic Commerce*, 1999, pp. 140–147.

[7] T. Mullen R. Jain and R. Hausman, “Analysis of Paris Metro pricing for QoS with a single service provider,” in *Proc. of Intnl. Workshop on QoS (IWQoS)*, June 2001, LNCS, 2001, Vol. 2092/2001, pp. 44–58.

[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,” in *Proc. of IEEE INFOCOM*, 2002, pp. 276–283.

[9] V. S. Borkar and D. Manjunath, “Charge-based control of diffserv-like queues,” *Automatica*, vol. 40, pp. 2043–2057, 2004.

[10] H. Mendelson and S. Whang, “Optimal incentive compatible priority pricing for the M/M/1 queue,” *Oper. Res.*, vol. 38, pp. 870–883, 1990.

[11] P. Dube and R. Jain, “N-player bertrand-cournot games in queues: Existence of equilibrium,” in *Proc. of the 46th Allerton Conf.*, Sept. 2008, pp. 491–498.

[12] G. J. van Houtum, *et al*, “Performance analysis of parallel identical machines with a generalized shortest queue arrival mechanism,” *OR Spektrum*, vol. 23, 2000.

[13] R. Tandra, N. Hemachandra, and D. Manjunath, “Join minimum cost queue for multiclass customers: Stability and performance bounds,” *Prob. in the Engg. and Inform. Sciences*, vol. 18, pp. 445–472, 2004.