The Concert Queueing Game: Processor Sharing Regime

Sandeep Juneja
Tata Institute of Fundamental Research, Mumbai
juneja@tifr.res.in

Tushar Raheja
Indian Institute of Technology Delhi
tushar@raheja.org

December 26, 2012

Abstract

We first survey the evolving literature on the concert/cafeteria queueing problem. This problem corresponds to determining the equilibrium arrival profile of non-cooperative customers selecting their arrival times to a queue where the service opens at a specified time. The customers are allowed to arrive before or after this time; they prefer to not wait and be served as soon as possible. This problem has a variety of queuing applications including how people queue at airport, movie theaters, passport offices, ration lines, etc. This also captures the settings where large computational jobs are sent to servers that open for service at a specified time. Substantial literature is devoted to studying the more tractable fluid version of this problem, that is, each customer is considered an infinitesimal particle, resulting in a non-atomic game between customers. This allows for explicit determination of the unique equilibrium arrival profile in many such settings as well as the associated socially optimal centralized solution. The knowledge of both then allows the computation of price of anarchy in the system. The literature thus far focuses on queues with the first come first serve service discipline. In this paper we again consider the fluid regime and extend the analysis to the case where the service discipline is processor sharing, or equivalently, random order service. The former is relevant in computational settings while the latter is a good approximation to settings where a customer is selected more or less at random by the server.

Keywords: Queueing Games, Nash Equilibrium, Processor Sharing, Random Order Service, Fluid Queues.

1 Introduction

In this paper we consider the concert or cafeteria queueing game of arrivals. The game involves a large finite population of non-cooperative customers arriving into a queueing system which starts service at a specified time. The customers choose their arrival time with the twin aim of minimizing the costs associated with the waiting in the queue and the time to finish service. The model was introduced by Juneja and Jain [7] and is motivated by familiar queues in cafeterias, movie theaters, bus-stops at bus route origin, airplane boarding, passport offices, banks, stores during black-Friday and similar large scale sales of electronics, etc.

Figure 1: Long queue example: Queue to purchase I-Pad2 in Beijing China

The model also finds application where computational jobs arrive to servers that become available at a specified time. The trade-off in the game can be summarized by considering the huge queues before a concert or a film where seats are not pre-assigned. If one goes early to occupy the best seats, one faces a longer queue, and when the queue is shorter later, the good seats are taken.

1.1 Literature Review

This problem was considered in [7] in the fluid framework, which can be seen as a limiting regime as the number of customers increases to infinity. Each person is a point in an interval that represents the total population. The cost structure of each customer is assumed to be linear and additive in waiting time and time to service. Customers are assumed to be homogeneous in that they have the same linear cost and share the same cost coefficients. The authors explicitly identify the unique Nash equilibrium arrival profile in this framework. The socially optimal or the social welfare solution (when a central planner schedules the service time of each customer) is straightforward in this setting and they show that the price of anarchy (PoA) equals 2. As is well known, PoA is defined as the ratio of the expected total cost incurred by all the customers under a worst cost Nash equilibrium to the total cost under a social welfare solution. Jain, Juneja and Shimkin [6], extend this analysis to multiple classes of customers, each with different cost parameters. They also show that a variety of modifications and what if analysis can be easily conducted in the fluid framework illustrating the tractability offered by this framework.

Juneja and Shimkin [9] extend the basic fluid model analysis to the finite population case. They show that equilibrium[1] in the finite setting has to be symmetric. This differs from the fluid setting where the unique arrival profile can also be attained by asymmetric customer behavior. Further, they identify the functional ordinary differential equation that each customer's arrival

[1]by equilibrium, we mean Nash equilibrium

distribution uniquely satisfies. One of their key contributions is to show that as the number of customers increases to infinity, the equilibrium arrival profile converges to the equilibrium solution of the fluid model. This thus builds credibility in the fluid analysis which is typically much more tractable.

An important assumption in [7] and [6] is that the total population is fixed. In [9] while they allow the population to be random, asymptotically they assume that the population converges to a fluid model with fixed population. This is typically not true in practice where the population arriving such queues maybe large but random. This problem of random arrival population was analyzed recently in the fluid regime by Juneja, Raheja and Shimkin [8]. They also show existence of a unique arrival profile where customers arrive uniformly over an interval. However, post this interval, the arrival density tapers off to zero with increasing time. Hence, customers have a higher arrival density in the beginning of the arrival's support than at the end. Interestingly, even in this setting PoA can be seen to equal 2 if one assumes that the central planner optimally scheduling the customers is aware of the distribution of the arriving customers but not its random realization.

Honnappa and Jain [5] extend the fluid concert queueing problem to a network of parallel queues where they again derive the unique equilibrium arrival profile. Haviv in [4] considers different modifications of the concert queue model in the finite as well as fluid setting.

Glazer and Hassin [2] were perhaps the first to consider strategic arrival decisions by customers arriving to a queue. They considered a framework where a total of Poisson distributed customers arrived at a queueing facility. The service times of customers was exponentially distributed and customers cost was linear in waiting time. They derived an ordinary differential equation whose solution gave the equilibrium arrival density of each customer (assumed to behave symmetrically). Many extensions of this basic model have since been considered. A comprehensive review of strategic decision models in queueing systems can be found in [3].

Bottleneck fluid models similar to the concert queue have been extensively studied in the transportation literature. Studying equilibrium patterns in road traffic was initiated by seminal papers of Wardrop [13] and Vickrey [12]. In Vickrey's model, also known as the morning commute problem, customers are again fluid particles and they have to decide at what time to leave for office separated via a bottleneck queue. They have a fixed preferred time: arriving too early has a penalty and so does lateness. Also see [11, 10] and references therein for further work in this area. Controlling equilibrium costs through information has been developed widely in Arnott and Lindsey [1] and other related papers by the authors.

1.2 Our Contributions

In this paper we review the analysis for the concert queueing game in the single class fluid regime and some related extensions. In addition, we consider this problem in the fluid setting when the service discipline is no longer first come first serve but is processor sharing or equivalently (in the fluid setting) random order service. Processor sharing corresponds to the server equally dividing its service amongst all the customers in the queue. This is relevant in many computational settings where the server rapidly time shares amongst customers present. For all practical purposes it is as if the server is equally dividing its service amongst all customers. Random order service could be prevalent in some computer queues where once a given job is completed the next one is selected randomly. It may also approximately model rowdy queues where the next person to be served is more-or-less randomly selected (not entirely unrealistic in India). In this setting we find that the equilibrium structure depends on model parameters. When customers put more weight on completing service early compared to waiting, the equilibrium

corresponds to everyone coming at time zero. On the other hand, if customers are more averse to waiting compared to completing the service early, the equilibrium profile corresponds to a point mass at time zero and customers arriving uniformly thereafter. The social welfare solution is easily seen in this case. It is similar to the case where the service is first come first serve. The PoA is less than 2 in the former case while it equals 2 in the latter.

The remaining paper is organized as follows: In Section 2 we develop the mathematical framework and review the equilibrium strategy for the deterministic population first come first serve (FCFS) setting. This is generalized to random fluid population setting in Section 3. In Section 4 we identify the unique equilibrium solution for the deterministic population setting where the service discipline is processor sharing. We end with a brief conclusion in Section 5.

2 Deterministic Population Fluid Model under FCFS

We assume that each customer is a point in an interval $[0, 1]$. Service starts at time zero, and continues thereafter at a constant rate $\mu > 0$. The costs incurred by each customer are taken to be linear and additive in waiting time and time to service.

If $(G_s(\cdot) : 0 \leq s \leq 1)$ denotes the collection of arrival profiles used by each customer (customer s samples its arrival time from distribution $G_s(\cdot)$) then let F denote the aggregate arrival profile where

$$F(t) = \int_s G_s(t)ds.$$

Note that due to the fluid nature of the customers, $F(t)$ denotes the deterministic amount of customers that arrive by time t. Let $W_F(t)$ be the waiting time of an arrival at time t in this scenario. The cost of an arrival at t to the serving facility is given by

$$C_F(t) = \alpha W_F(t) + \beta(t + W_F(t))$$

where $t + W_F(t)$ is the time to service of a customer who arrives at time t. Here $\alpha > 0$ is the unit cost of waiting time in the system and $\beta > 0$ is the unit cost of time to service. Without loss of generality we assume that $\alpha + \beta = 1$; in particular, $0 < \alpha, \beta < 1$. Note that due to fluid analysis, this cost is the same for each arrival at time t and depends only on the aggregate profile F.

Let $Q_F(t)$ denote the queue size at time t. Then if F does not have a jump at time t, under FCFS service discipline,

$$W_F(t) = Q_F(t)/\mu + \max\{0, -t\}.$$

The cost of a customer who selects her arrival time by sampling from probability distribution H (recalling that $\alpha + \beta = 1$) is

$$C_{H,F} = \int_{-\infty}^{\infty} [W_F(t) + \beta t] \; dH(t).$$

Definition 1 *A multi-strategy $(G_s(\cdot) : 0 \leq s \leq 1)$ with aggregate profile F is in Nash equilibrium if no customer can unilaterally improve its cost by changing its strategy. That is,*

$$C_{G_s,F} \leq C_{H,F}$$

for all H, for each $s \in [0, 1]$.

It is easy to see that this corresponds to existence of an arrival profile F such that there exists a set $\mathcal{T}'$ of F measure 1 and a constant c such that

$$C_F(t) \geq c \tag{1}$$

for all t, and

$$C_F(t) = c \tag{2}$$

for all $t \in \mathcal{T}'$.

To see the equivalence of the two criteria for equilibrium note that if there exists an arrival profile F and set $\mathcal{T}'$ such that (1) and (2) hold, then, setting each G_s to F it is easy to see that the resulting multi-strategy is in equilibrium. Alternatively, suppose we have a multi strategy that is in equilibrium and the resultant F does not satisfy (1) and (2), then there must exist a set of positive F measure, call it A, where the cost $C_F(t)$ is higher compared to at another time, call it, s. Then, a customer that has a positive mass at A can improve its cost by putting some of that mass at s, thereby providing the desired contradiction.

Note that we haven't ruled out the fact that given an F that satisfies (1) and (2), there exist multiple multi-strategies that are in equilibrium. Indeed, the latter is true. Later we give some examples.

2.1 Existence and Uniqueness of Equilibrium Profile

Now we argue that in our game there is a unique F that satisfies (1) and (2) for a unique closed interval, $\mathcal{T}$, the closure of $\mathcal{T}'$. We do this by fathoming the set of possible equilibrium profiles till only one remains.

We make a number of observations regarding an equilibrium profile F:

1. There are no point masses in F. If there were, then a customer arriving just before such a point incurs less waiting and is served earlier than any arrival at that point.

2. The cost under F at each t must be at least β/μ. This is true since the last customer must be served at time $\geq 1/\mu$ (since if the server serves at full rate μ it needs time $1/\mu$ to serve all the customers).

3. Under F, the server works at a full rate μ until the last customer is served at time $1/\mu$. This follows as if instead under F there were a time t before time $1/\mu$ where the server has spare capacity, then $C_F(t) < \beta/\mu$, a contradiction!

4. The above observation also means that the end-point t_e of support of F is less than or equal to $1/\mu$. (Recall that a support of F is the smallest closed set of F measure 1. It is the closure $\mathcal{T}$ of $\mathcal{T}'$.)

5. $t_e = 1/\mu$. To see this suppose that $t_e < 1/\mu$. Then there exists a queue at time t_e so that $C_F(t_e) > \beta/\mu$. However, if a customer arrives at time $1/\mu$, it does not encounter a queue and its cost is β/μ. It follows that $t_e = 1/\mu$.

6. It then follows that $C_F(1/\mu) = \beta/\mu$ and that the cost incurred by an arrival at any point in the support of F must be β/μ. It of course must be at least as high elsewhere.

7. It also follows from 3 that $F(t) \geq \mu t$ for $0 \leq t \leq 1/\mu$.

8. Hence, $W_F(t) = F(t)/\mu - t$ for $t \leq 1/\mu$ and $W_F(t) = 0$ otherwise. This expression is obvious for $t < 0$ since customer arriving at time $t < 0$ has to wait for $-t$ for the server to start serving, and it has to wait $F(t)/\mu$ for the customers that arrived earlier to get served. For $t > 0$, this follows from observation 3 above.

9. Hence, the cost function $C_F(t)$ equals

$$\beta(t + W_F(t)) + \alpha W_F(t) = F(t)/\mu - \alpha t,$$

for $t \leq 1/\mu$.

10. Furthermore, this equals β/μ at $1/\mu$ and along the support of F. It follows that F does not have any gap in its support so that

$$F(t) = \frac{(t - t_b)}{(t_e - t_b)},$$

where $t_b = -\beta/(\alpha\mu)$ is the beginning of the support of F. This is the time at which customers start arriving at the queue. $\mathcal{T} = [t_b, t_e]$. $F(t) = 0$ for $t < t_b$ and $F(t) = 1$ for $t > t_e$.

Above arguments limit by necessity, the equilibrium arrival profile to a single F. The fact that this F is indeed an equilibrium profile is easily checked. The cost of an arrival in the interval $[t_b, t_e]$ is constant β/μ while it is higher for each $t < t_b$ or $> t_e$.

It therefore follows that the arrival profile is unique and is uniformly distributed between $[-\beta/(\alpha\mu), 1/\mu]$. Note the density of the arrival profile is constant and equals

$$\frac{\alpha}{\alpha + \beta}\mu = \alpha\mu,$$

along the interval $[-\beta/(\alpha\mu), 1/\mu]$. Note that this arrival profile can be obtained in many different ways. For instance, one way involves each arrival selecting her arrival time uniformly along the interval $[-\beta/(\alpha\mu), 1/\mu]$. Alternatively, half the population may select their arrival time uniformly from first half of this interval and the remaining half may select their arrival time uniformly from the remaining half of the interval. Another alternative is that the customers arrive deterministically along this interval at a constant rate $\alpha\mu$.

Figure 2 shows the queue length process and the arrival profile under equilibrium. Note that each customer incurs a cost β/μ under this equilibrium. Hence the total cost to all customers (or the social equilibrium cost) is also β/μ.

The social optimal solution is easily seen in this setting. There will be no queue as each customer can be scheduled to arrive at the instant his service starts so there is no waiting. Clearly, the server has to serve at the fastest possible rate between interval $[0, 1/\mu]$ otherwise the cost can be improved by scheduling customers during server unutilized time. It follows that in the social optimal solution customers arrive at a uniform rate μ between $[0, 1/\mu]$ and they are served at rate μ during this time so there is no queue.

Average service completion time then is $1/(2\mu)$ so that the total cost of all customers equals $\beta/(2\mu)$. Hence, the price of anarchy equals 2.

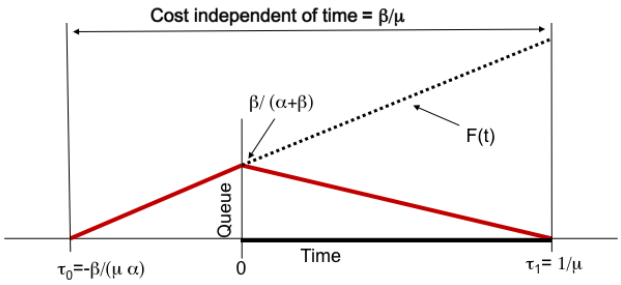

Figure 2: The queue length process and the arrival profile in equilibrium

2.2 Multi Class Setting

In [6] they consider multi-class customers in the sense that while each class still has linear costs as above, the cost coefficients for class i are allowed to be different and may be given by (α_i, β_i). They show that in equilibrium the different classes arrival times separate into contiguous, disjoint intervals and they arrive in descending order of β_i/α_i (when there is a tie in this ratio, those classes can be thought of as a single class). The last customer to arrive again has zero wait. The arrival density function of each class is given by

$$\frac{\alpha_i}{\alpha_i + \beta_i}\mu.$$

This ensures that the cost is constant along the support of arrival profile of each class. It can be checked that the specified ordering of the classes ensures that any customer coming at a time different from the support of its arrival profile incurs a higher cost. Note that the customers arrive slowly at first and then at a faster rate. Figure 3 illustrates the queue length process in the multi-class settings.

Under social optimal solution again there is no queueing. It is easy to see that customers come segregated by class at rate μ each. Class with higher β comes earlier to minimize costs. It can be seen that PoA may no longer equal 2 when two or more classes are involved. [6] also develop tight upper and lower bounds on the PoA.

3 Uncertainty in Arrival Volume

This case was considered in [8]. They again consider a fluid model where the total single class arrival population is given by $[0, \Lambda]$ where Λ is a random variable with distribution function G_0 which is known to all customers. The queue discipline is again FCFS. They show that the resulting equilibrium profile is no longer uniform. The shape of F is now concave and the equilibrium cost is higher compared to the deterministic case when the total arrival volume equals $E\Lambda$. The socially optimal profile remains uniform, although there may be queueing when Λ is sufficiently large. We now further detail this discussion.

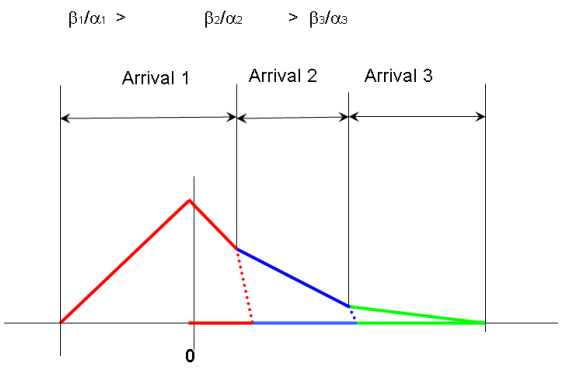

Figure 3: The queue length process for multi-class customers in equilibrium

First note that each arriving customer sees a distribution of remaining arriving customers that differs from G_0. Specifically, she sees the distribution G given by

$$dG(\lambda) = \frac{\lambda dG_0(\lambda)}{\int_0^\infty \lambda dG_0(\lambda)}$$

This length biased distribution captures the fact that a particular arrival is more likely to arrive in a larger volume of arrivals as compared to when a lesser volume of customers arrive (see, e.g., [6] for further explanation).

In the analysis, one now needs to condition on the realization Λ. Thus, $W_{\Lambda,F}(t)$ denotes the waiting time of an arrival at time t when the overall arrival amount equals Λ and the distribution followed by each arrival is F. The cost incurred by this arrival equals

$$C_{\Lambda,F}(t) = \alpha W_{\Lambda,F}(t) + \beta(t + W_{\Lambda,F}(t))$$

The results are derived in much the same way as the deterministic case. A point mass in the equilibrium profile F is similarly ruled out. It can be shown that the support of F equals an interval $[t_b, t_e]$ with $-\infty < t_b < 0$ and $0 < t_e < \infty$.

Let

$$T_\Lambda = \inf\{t \geq 0 : \Lambda F(t) < \mu t\} \tag{3}$$

denote the first time after zero when the server starts to serve at less than full rate μ when Λ customers arrive. The equilibrium analysis for this case begins by proving that queue does not rebuild after time T_Λ. This then implies that

$$W_{\Lambda,F}(t) = \Lambda F(t)/\mu - t$$

for $t \leq T_\Lambda$, and $W_{\Lambda,F}(t) = 0$ otherwise.

The expected cost for an arrival at time t can be written as,

$$EC_F(t) = \int_{\frac{\mu t}{F(t)}}^\infty \left(\lambda \frac{F(t)}{\mu} - t\right) dG(\lambda) + \beta t.$$

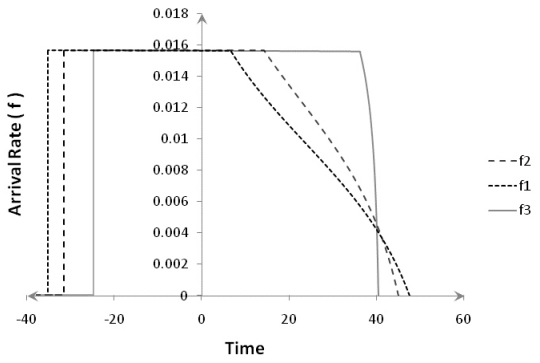

Figure 4: G is uniformly distributed: f_1 - within $[50, 350]$; f_2 - within $[100, 300]$; and f_3: - within $[190, 210]$. Here, $\beta = 3/8, \mu = 5$.

For the cost to be invariant with time, equilibrium profile is found by simply equating its derivative to zero to get

$$F'(t) = \mu \frac{(\bar{G}(\frac{\mu t}{F(t)}) - \beta)}{\int_{\frac{\mu t}{F(t)}}^{\lambda_h} \lambda dG(\lambda)} \tag{4}$$

along the interval (t_b, t_e).

Define

$$\lambda^* = \inf\{\lambda : G(\lambda) \geq \alpha\} \tag{5}$$

(we are again assuming that $\alpha + \beta = 1$). Evidently, $G(\lambda^*) = \alpha$, unless λ^* is a discontinuity point of G. In the former case, it is easy to argue that $F'(t_e^-) > 0$ implies that the cost improves for an arrival just after time t_e, which contradicts F being an equilibrium profile, so we have $F'(t_e^-) = 0$. This implies that $t_e = \frac{\lambda^*}{\mu}$. This result can be seen to hold even when λ^* is a discontinuity point of G.

Let $\lambda_l \geq 0$ denote the left limit of support of G, corresponding to the minimal possible arrival volume (minimum value of Λ). Recall the definition of T_Λ from (3). Then, till T_{λ_l} the server serves at full rate for any value of Λ. Thereafter, there exist Λ values for which the server serves at less than full capacity.

The equilibrium cost c_e equals $EC_F(t_e)$ so that

$$c_e = \frac{1}{\mu} \int_{\lambda^*}^{\infty} \bar{G}(\lambda) d\lambda + \beta \frac{\lambda^*}{\mu}.$$

The time of first arrival t_b then equals $-\frac{c_e}{\alpha}$.

By differentiating (4) it can be seen that F is concave.

In Figure 4, we plot the arrival profiles when the distribution G of Λ is uniform between λ_l and λ_h for three sets of values of λ_l and λ_h. Observe that for mean $(\lambda_h + \lambda_l)/2$ fixed, the equilibrium cost increases with increasing variance. This also leads to increase in deviation from uniform distribution in the arrival profile.

For determining the socially optimal profile, for fair comparison, we consider a central scheduler who prescribes $F_s(t)$ in advance without knowing the realization Λ. Thus, $\Lambda F_s(t)$ denotes the volume of arrivals by time t. Ascertaining the socially optimal arrival profile is no longer straightforward and involves solving an intricate calculus of variations problem (see [8]). It can be seen that F_s is a uniform distribution with density

$$F_s'(t) = \frac{\mu}{\lambda^*}, \quad 0 \leq t \leq t_e = \frac{\lambda^*}{\mu}$$

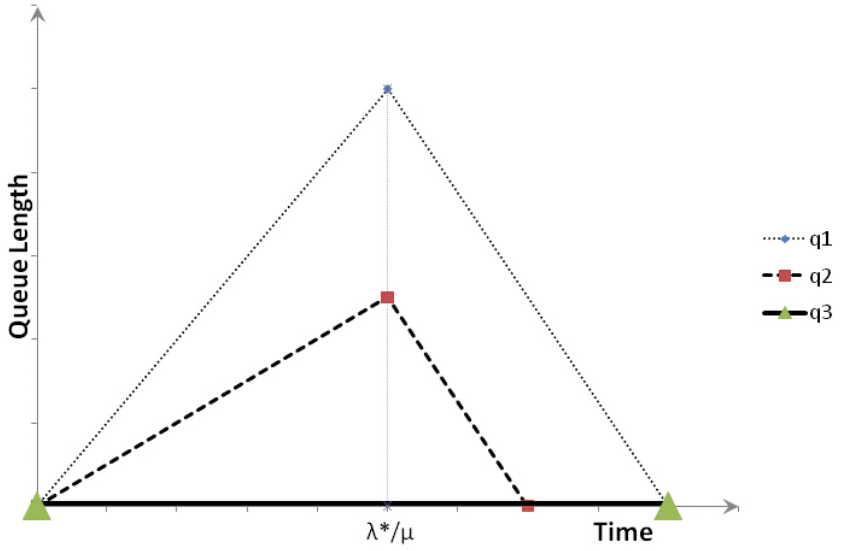

Figure 5: If $\Lambda \leq \lambda^*$, then no one waits. If $\Lambda > \lambda^*$, then everyone waits.

where λ^* is defined in (5). Remarkably, the PoA remains 2 in this case as well. Figure 5 illustrates the queue length process under this socially optimal solution. Thus, if Λ is less than or equal to λ^* there is no queueing (q3 profile in Figure 5), if Λ is greater than λ^* then the queue grows tile time $t_e = \lambda^*/\mu$ and thereafter it empties at rate μ (q1 and q2 profile in Figure 5).

4 Processor sharing or service in random order

In this section we introduce a new variant of the concert game where customers instead of being served in the FCFS discipline, are served as in the fluid limit of the processor sharing regime, or equivalently, the random order service regime. Essentially, we assume that if at time t, $Q(t) > 0$ denotes the volume of customers in the system, then the probability that a customer in the queue gets served in the next infinitesimal interval Δt is given by

$$\frac{\mu \Delta t}{Q(t)} + o(\Delta t).$$

The remaining assumptions are same as for the FCFS game. Service starts at time zero, and continues at a constant rate $\mu > 0$ as long as there are sufficient customers in the queue. The costs incurred by each customer are taken to be linear and additive in waiting time and time to service with parameters α and β. The volume of arrivals to the queue on any day, Λ, is again taken to be 1. We again set $\alpha + \beta = 1$ for notational convenience.

Note that in this service regime, in equilibrium, customers have no incentive to come before time zero as then they can improve their cost by coming at time 0. It turns out that the equilibrium profile has an interesting solution that depends on the parameters α and β. Specifically, when $\alpha \leq \beta$ then in equilibrium all customers arrive at time 0. When $\alpha > \beta$, the equilibrium profile is more interesting. At time zero, β/α amount of customers arrive. The remaining $1 - \beta/\alpha$ amount of customers arrive uniformly between times $[0, 1/\mu]$. We now illustrate this analysis.

4.1 $\alpha \leq \beta$

In this case if all customers come at time zero then their average wait and time to service both equal $1/(2\mu)$ so their average cost is $(\alpha+\beta)/(2\mu) = 1/(2\mu)$. We now argue that this is indeed an equilibrium strategy.

Suppose that all other customers arrive at time zero and one customer comes at time $0 < \delta \leq 1/\mu$. At time δ there remain $1 - \mu\delta$ customers in queue. So this customer's average waiting time is

$$(1 - \mu\delta)/(2\mu).$$

It will complete its service at an average time of

$$\delta + (1 - \mu\delta)/(2\mu).$$

Therefore, its average cost is

$$\alpha(1 - \mu\delta)/(2\mu) + \beta(\delta + (1 - \mu\delta)/(2\mu)),$$

which equals

$$1/(2\mu) + \delta(\beta - \alpha)/2.$$

Since, $\beta \geq \alpha$ this customer has no incentive to come after time zero, so that the strategy of all customers arriving at time zero constitutes an equilibrium.

The social optimal in this case remains the same as in FCFS, namely that customers arrive at rate μ during the interval $[0, 1/\mu]$ so that there is no queue and the total social optimal cost equals $\beta/(2\mu)$. It then follows that the price of anarchy equals $1/\beta$. This, surprisingly, is less than two when $\beta > \alpha$!

4.2 $\alpha > \beta$

From the arguments given in the previous section, it follows that for $\alpha > \beta$ everyone coming at time zero is no longer an equilibrium.

Suppose that in equilibrium the aggregate profile is given by F and its support by $\mathcal{T}$ (recall that support is the smallest closed set of F measure 1). It can be easily seen that again the server must serve at full rate till time $1/\mu$, in particular $F(t) - \mu t > 0$ for $0 \leq t < 1/\mu$ (else a customer can improve cost by coming at a time where the queue is empty). This also implies that $F(1/\mu) = 1$ so that $\mathcal{T} \subset [0, 1/\mu]$. Further, the queue length $Q(t) = F(t) - \mu t$ for $0 \leq t \leq 1/\mu$. In particular, the effective service rate for any customer present at time t, call it $\mu(t)$, for $0 < t < 1/\mu$ satisfies

$$\mu(t) = \frac{\mu}{F(t) - \mu t}.$$

Let

$$m_t(s) = \int_t^s \mu(y)dy.$$

For any **customer** arriving at time t, the time to departure is the time to first event of a non-homogeneous Poisson process with rate $\mu(t)$. It follows that the density function that an arrival at time t departs at time s, call it $h_t(s)$, may be seen to equal

$$h_t(s) = \frac{\mu}{F(s) - \mu s} \exp(-m_t(s)).$$

To see that this is a valid density function, we need to show that

$$\int_t^{1/\mu} h_t(s)ds = 1$$

for all $0 < t < 1/\mu$. Note that LHS above equals

$$\exp(-m_t(t)) - \exp(-m_t(1/\mu)) = 1 - \exp(-m_t(1/\mu)).$$

Thus we need to show that $m_t(1/\mu)) = \infty$. To see this, note that

$$m_t(1/\mu)) = \int_t^{1/\mu} \frac{\mu}{F(s) - \mu s} ds \geq \int_t^{1/\mu} \frac{\mu}{1 - \mu s} ds.$$

Since, the RHS equals infinity, the result follows.

Hence, the expected waiting time of a customer that arrives at time t can be expressed as

$$EW_F(t) = \mu \int_t^{\frac{1}{\mu}} \frac{s - t}{F(s) - \mu s} \exp\left(-m_t(s)\right) ds. \tag{6}$$

Note that $EW_F(t)$ is a continuous and differentiable function of t.

The expected cost (taking as before $\alpha + \beta = 1$),

$$C_F(t) = EW_F(t) + \beta t.$$

Thus, $C_F(t)$ is a continuous and differentiable function of t and is constant on the support $\mathcal{T}$ of F. Let c_e denote this constant cost. Hence, the derivative $C_F'(t)$ in the interior of $\mathcal{T}$ is zero. That is, for $t \in \mathcal{T}^o$,

$$\frac{dEW_F(t)}{dt} + \beta = 0,$$

or

$$\mu \int_t^{\frac{1}{\mu}} \frac{1}{F(s) - \mu s} \left(\frac{(s-t)\mu}{F(t) - \mu t} - 1 \right) \exp\left(- \int_t^s \mu \frac{1}{F(r) - \mu r} dr \right) ds + \beta = 0$$

which is rearranged as,

$$\frac{\mu EW_F(t)}{F(t) - \mu t} - \int_t^{\frac{1}{\mu}} h_t(s)ds + \beta = \frac{\mu EW_F(t)}{F(t) - \mu t} - 1 + \beta = 0 \tag{7}$$

Recall that $EW_F(t) = c_e - \beta t$ and substitute this in (7). Thus, equilibrium profile must satisfy

$$F(t) - \mu t = \mu \frac{c_e - \beta t}{\alpha} \tag{8}$$

for $t \in \mathcal{T}^o$, or

$$F(t) = \frac{\mu}{\alpha}((\alpha - \beta)t + c_e)$$

for $t \in \mathcal{T}^o$ so that $\alpha \geq \beta$ is necessary for $\mathcal{T}^o$ to be non-empty. It also follows that if $\alpha < \beta$ then $\mathcal{T}^o$ is empty. This is a key observation in establishing the uniqueness of the equilibrium profile proposed earlier for $\alpha < \beta$. The case $\alpha = \beta$ requires further arguments.

Number of observations can be seen to follow from (8):

1. F cannot have a jump at at any point in $\mathcal{T}^o$.

2. F cannot have a jump at at any point in $\mathcal{T}$. To see this, suppose that there exist $0 \leq t_1 < t_2 < t_3 < t_4 \leq 1/\mu$ such that $[t_1, t_2] \cup [t_3, t_4] \in \mathcal{T}$ and there is a point mass either at t_2 or at t_3. We argue against these possibilities. (In our analysis, as in [9] and other related references, we have have limited ourselves to $\mathcal{T}$ that can be represented as a finite union of intervals along any bounded set).

   From (8) it follows that

$$F(t_2^-) - \mu t_2^- = \mu \frac{c_e - \beta t_2^-}{\alpha} = \mu \frac{c_e - \beta t_2}{\alpha}$$

   where the statement regarding t^- corresponds to statement regarding $t - \epsilon$ for $\epsilon > 0$ in the limit as ϵ decreases to zero (similarly, t^+).

   Now, $C_F'(t_2^+) \geq 0$ (note that $C_F(t)$ is differentiable). Carrying the analysis as above, noting by continuity that $EW_F(t_2^+) = EW_F(t_2^-) = c_e - \beta t_2$, it follows that

$$F(t_2^+) - \mu t_2^+ \leq \mu \frac{c_e - \beta t_2^+}{\alpha} = \mu \frac{c_e - \beta t_2}{\alpha},$$

   so that $F(t_2^+) = F(t_2^-)$ and there is no jump at t_2. Similarly, we can show that $F(t_3^-) = F(t_3^+)$.

3. Next, it is easy to see that gaps such as (t_2, t_3) in $\mathcal{T}$ cannot exist as that would imply that $F(t_2) = F(t_3)$ leading to a contradiction since (8) holds at t_2^- and at t_3^+. This implies that the support of F must be an interval, call it, $[t_b, t_e]$.

4. Furthermore, the derivative of F must exist along (t_b, t_e) and $F'(t) = \mu \frac{\alpha - \beta}{\alpha}$ for $t \in (t_b, t_e)$.

5. It can be easily seen that $t_b = 0$ for if $t_b > 0$ then a customer can improve its cost by coming at time zero.

6. Furthermore, if $t_e < 1/\mu$ then one argument illustrating that this cannot be an equilibrium is as follows: We show later that if $t_e = 1/\mu$ then $c_e = \beta/\mu$. This implies that if $t_e < 1/\mu$ then, all customers arrive earlier than they would if $t_e = 1/\mu$, implying that, since each arrival sees relatively more people ahead of it when it joins, it incurs a higher cost so that in this case $c_e > \beta/\mu$. This cannot be an equilibrium because a customer can then improve its cost by simply arriving at $1/\mu$ to an empty queue. Thus, $t_e = 1/\mu$.

Now we show that the profile necessitated by above discussion is indeed an equilibrium profile where each customer incurs an expected cost of β/μ.

To see this, note that the candidate profile is given by

$$F(t) = \beta/\alpha + (1 - \beta/\alpha)\mu t$$

for $0 \leq t \leq 1/\mu$. $F(t) = 0$ for $t < 0$.

Then, for $0 \leq t < 1/\mu$,

$$m_t(s) = \int_t^s \frac{\mu}{F(r) - \mu r} dr = \frac{\alpha}{\beta} \int_t^s \frac{\mu}{(1 - \mu r)} dr.$$

This simplifies to equal

$$\frac{\alpha}{\beta} \log \left(\frac{1 - \mu t}{1 - \mu s} \right).$$

Then, for $0 \leq t < 1/\mu$

$$EW_F(t) = \mu\frac{\alpha}{\beta}\int_t^{1/\mu} \frac{s-t}{1-\mu s}\exp\left(-\frac{\alpha}{\beta}\log\left(\frac{1-\mu t}{1-\mu s}\right)\right)ds.$$

After some calculus it can be seen that

$$EW_F(t) = (1/\mu - t)\beta$$

so that $c_e = \beta/\mu$. In particular the proposed profile is indeed a unique equilibrium profile.

The social optimal does not depend on the service order and remains the same. The PoA, therefore remains 2 when $\alpha > \beta$.

5 Conclusion

In this paper we reviewed the evolving literature on ascertaining the equilibrium arrival profiles in a concert queueing game. We considered in some depth the basic model where a deterministic number of customers with linear and homogeneous costs in waiting and time to service want to join a queueing service that opens at a specified time and serves FCFS at a fixed rate. We discussed some variations including when the arrivals are no longer homogeneous but belong to multiple classes. We also reviewed the single class case where the arrivals number is a random quantity.

Our chief contribution was to develop the equilibrium profile in the setting where the service discipline is no longer FCFS but is instead processor sharing or random order service. We found that in this case the equilibrium profile substantially differs from the FCFS case and can be parameter dependent. In particular, for certain parameter settings the total cost of the equilibrium (summing over all customers) is actually an improvement over the FCFS case while in the remaining parametric settings the two have the same total equilibrium cost.

References

[1] R. Arnott, A. Palma and R. Lindsey. 1999. Information and time-of-usage decisions in the bottleneck model with stochastic capacity and demand *European Economic Review* **43**, 525-548.

[2] A. Glazer and R. Hassin. 1983. '?/M/1: On the equilibrum distribution of user arrivals, *Eur. J. Oper. Res.* 13:146-150.

[3] R. Hassin and M. Haviv. 2003. *To Queue or Not to Queue.* Kluwer Academic Publishers.

[4] M. Haviv. 2010. When to arrive at a queue with tardiness costs? Preprint.

[5] H. Honnapa and R. Jain. 2010. Strategic arrivals into queueing networks. *Proc. 48th Annual Allerton Conference*, Illinois, Oct. 2010, pp. 820-827.

[6] R. Jain, S. Juneja and N. Shimkin. 2011. The Concert Queuing problem: To wait or to be late *Discrete Event Dyn. Syst.* **21**, 103-138.

[7] S. Juneja and R. Jain. 2009. The Concert/Cafeteria Queuing Problem: A Game of Arrivals, *in Proc. ValueTools'09 Fourth ICST/ACM Fourth International Conference on Performance Evaluation Methodologies and Tools*, Pisa, Italy.

[8] S. Juneja, T. Raheja and N. Shimkin. 2012. The Concert Queuing Game with Random Arrivals Volume, *in Proc. ValueTools'12 6th International Conference on Performance Evaluation Methodologies and Tools*. 317-325.

[9] S. Juneja and N. Shimkin. 2012. The Concert Queueing Game: Strategic arrivals with waiting and tardiness costs, *Queueing Systems*. DOI 10.1007/s11134-012-9329-3

[10] R. Lindley. 2004. Existence, uniqueness, and trip cost function properties of user equilibrium in the bottleneck model with multiple user classes, *Transport. Sci.* 38(3):293-314.

[11] G.F. Newell. 1987. The morning commute for nonidentical travellers, *Transport. Sci.* 21(2):74-88.

[12] W. S. Vickrey. 1969. Congestion Theory and Transport Investment *The American Economic Review* **59**, 251-260.

[13] J. G. Wardrop. 1952. Some Theoretical Aspects of Road Traffic Research. *Proc. Inst. Civil Engineers*, Part 2, Vol. 1, 325-378.