International Game Theory Review
Vol. 17, No. 2 (2015) 1540012 (15 pages)

DOI: 10.1142/S0219198915400125

World Scientific
www.worldscientific.com

The Concert Queueing Game: Fluid Regime with Random Order Service

Sandeep Juneja

School of Technology and Computer Science
Tata Institute of Fundamental Research
HB Road, Colaba, Mumbai-400 005, India
juneja@tifr.res.in

Tushar Raheja

Department of Mechanical Engineering
Indian Institute of Technology Delhi
Hauz Khas, New Delhi-110 016, India
tushar@raheja.org

Received 29 December 2012
Revised 28 October 2013
Accepted 12 September 2014
Published 19 March 2015

The concert queueing 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. 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 (PoA) in the system. The literature thus far focuses on queues with the first come first serve (FCFS) service discipline. In this paper, we again consider the fluid regime and extend the analysis to the case where the service discipline is random order service (ROS). This is equivalent to the practically equally important processor sharing regime when the service times are exponential. The latter is relevant in computational settings while the former is a good approximation to settings where a customer is selected more or less at random by the server.

Keywords: Queueing games; Nash equilibrium; random order service; processor sharing; fluid queues.

Subject Classification: 91A10, 91A13, 91A80

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. 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 [2009] 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.

This 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*

We first review the evolving literature on the concert queueing problem and then review closely related research.

Juneja and Jain [2009] considered this problem 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 it can be seen 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 *et al.* [2011], 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* are easily conducted in the fluid framework, thus illustrating the tractability offered by this framework.

Juneja and Shimkin [2012] extend the basic fluid model analysis to the finite population case. They show that equilibrium[a] 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 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

[a]By equilibrium, we mean Nash equilibrium.

solution of the fluid model. This thus builds credibility in the fluid analysis which is typically much more tractable.

An important assumption in Juneja and Jain [2009] and Jain *et al.* [2011] is that the total population is fixed. In Juneja and Shimkin [2012], 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 to such queues maybe large but random. This problem of random arrival population was analyzed in the fluid regime by Juneja *et al.* [2012]. 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, the PoA can be seen to equal 2 if one assumes that the central planner is aware of the distribution of the arriving customers but not its random realization.

Honnapa and Jain [2010] extend the fluid concert queueing problem to a network of parallel queues where they again derive the unique equilibrium arrival profile. Haviv [2013] considers different modifications of the concert queue model in the finite as well as fluid setting. These include cases where the arrivals are not allowed to queue before the server starts service and they are allowed to enter a queue up to a specified time.

Glazer and Hassin [1983] were the first to consider strategic arrival time 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 were assumed to be exponentially distributed and customers cost was linear in waiting time. The server starts service at a specified 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. Hassin and Kleiner [2010] consider a similar queue with the contingency that customers are not allowed to queue before the server starts service so that amongst those who arrive at that time, customers are randomly selected for service. Thereafter first-come-first-serve (FCFS) rule is followed.

A comprehensive review of strategic decision models in queueing systems can be found in Hassin and Haviv [2003].

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 [1952] and Vickrey [1969]. 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 Newell [1987]; Lindsey [2004] and the references therein for further work in this area. Controlling equilibrium costs through

information has been developed in Arnott *et al.* [1999] and other related papers by the authors.

1.2. *Our contributions*

In this paper, we first briefly review the analysis for the concert queueing game in the single class fluid regime when the service discipline is first come first served. We then consider this problem when the service discipline is no longer first come first serve but is random order service (ROS). As we later discuss, this case also corresponds to the limiting fluid limit of the queue where the service discipline is processor sharing and the service times are independent and identically exponentially distributed. As is well known, 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. ROS 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 framework, we find that the equilibrium structure depends on the 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. We also analyze our queueing model under the contingency that the arrivals are restricted to come before a specified time T. This turns out to be a straightforward generalization of our results in the unrestricted case. Haviv [2013] finds similar PoA behavior in the fluid models under the first-come first server discipline with restriction on customer arrival times.

The remaining paper is organized as follows: In Sec. 2, we develop the mathematical framework and review the equilibrium strategy for the FCFS setting. In Sec. 3, we identify the unique equilibrium solution when the service discipline is ROS and evaluate its PoA. We also discuss the equilibrium solution for this queue under a upper time restriction on arrival times. We end with a brief conclusion and some directions for further work in Sec. 4.

2. Fluid Model under FCFS

We first discuss the basic framework that is useful for both FCFS and ROS. Later in Sec. 2.1, we specialize to FCFS service discipline.

Assume that each customer is a point in an interval $[0, \Lambda]$. 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 in time to service.

If $(G_s(\cdot) : 0 \leq s \leq \Lambda)$ denotes the collection of arrival profiles used by each customer (customer s samples her 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 of service completion 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. 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 is

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

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

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

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

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 Λ 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 $t \in \mathcal{T}'$. Furthermore, such a $\mathcal{T}'$ is minimal in the sense that its closure $\mathcal{T}$ is the support of F (recall that a support of F is the smallest closed set of F measure Λ).

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/Λ, 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 her cost by putting some of that mass at s, thereby providing the desired contradiction.

Note that we have not 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*

We now review the results in Juneja and Jain [2009], Jain *et al.* [2011]. Specifically, we show that under the FCFS service discipline, in the above 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. The arguments provided are simpler than those in Juneja and Jain [2009] and Jain *et al.* [2011].

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 $\beta\Lambda/\mu$. This is true since the last customer must be served at time $\geq \Lambda/\mu$ (since if the server serves at full rate μ it needs time Λ/μ to serve all the customers).
(3) Under F, the server works at a full rate μ until the last customer is served at time Λ/μ. Furthermore, there is a positive queue at each time $t \in [0, \Lambda/\mu)$. For otherwise if queue was empty at anytime $t \in [0, \Lambda/\mu)$, then $C_F(t) < \beta\Lambda/\mu$, a contradiction!
(4) The above observation also means that the end-point t_e of support of F is less than or equal to Λ/μ.
(5) $t_e = \Lambda/\mu$. To see this suppose that $t_e < \Lambda/\mu$. Then there exists a queue at time t_e so that $C_F(t_e) > \beta\Lambda/\mu$. However, if a customer arrives at time Λ/μ, it does not encounter a queue and its cost is $\beta\Lambda/\mu$. It follows that $t_e = \Lambda/\mu$.
(6) It then follows that $C_F(\Lambda/\mu) = \beta\Lambda/\mu$ and that the cost incurred by an arrival at any point in the support of F must be $\beta\Lambda/\mu$. It of course must be at least as high elsewhere.
(7) It also follows from Observation 3 that $F(t) > \mu t$ for $0 \leq t < \Lambda/\mu$.
(8) Hence, $W_F(t) = F(t)/\mu - t$ for $t < \Lambda/\mu$ 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

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

for $t \leq \Lambda/\mu$.

(10) Furthermore, this equals $\beta\Lambda/\mu$ at Λ/μ 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 $t_b = -\beta\Lambda/(\alpha\mu)$ 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 $F(t) = \Lambda$ 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 $\beta\Lambda/\mu$ 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 $[-\beta\Lambda/(\alpha\mu), \Lambda/\mu]$. Note the density of the arrival profile is constant and equals

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

along the interval $[-\beta\Lambda/(\alpha\mu), \Lambda/\mu]$.

Also 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 $[-\beta\Lambda/(\alpha\mu), \Lambda/\mu]$. 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 $\frac{\alpha}{\alpha+\beta}\mu$.

Figure 1 shows the queue length process and the arrival profile under equilibrium (here $\Lambda = 1$). 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 $[0, \Lambda/\mu]$ 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 $[0, \Lambda/\mu]$ and they are served at rate μ during this time so there is no queue.

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

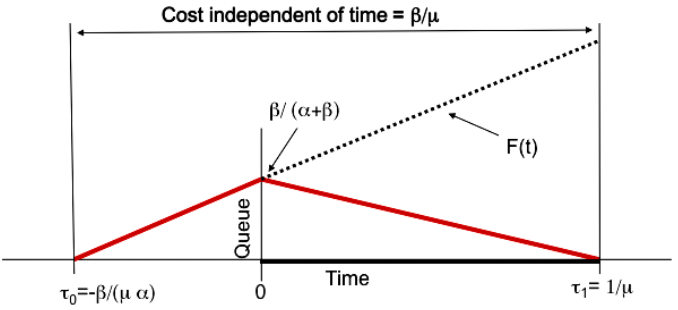

Fig. 1. Long queue example: Queue to purchase I-Pad2 in Beijing, China.

3. Random Order Service

In this section, we introduce a new variant of the concert game where customers instead of being served in the FCFS discipline, are served in the fluid limit of the ROS 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

$$\frac{\mu \Delta t}{Q(t)} + o(\Delta t). \tag{3}$$

The remaining assumptions are the 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.

Remark 1 (Processor Sharing when service times are exponential). It is important to note that (3) holds for a limiting processor sharing fluid model when the service times are independent and identically exponentially distributed, so that our analysis is valid for this regime as well.

To see this simply, suppose that at time nt the system has cn ($c > 0$) customers in the system and service requirement of each customer is independent and exponentially distributed with rate μ. Then, under the processor sharing regime, each customer at time t gets $1/(cn)$ of the server capacity and due to the independent exponentially distributed service times, the probability that a given customer in queue departs between time nt and $n(t + \Delta t)$ is $\frac{\mu}{c}\Delta t + o(\Delta t)$. Furthermore, the probability that there is a system departure between time nt and $n(t + \Delta t)$ is $n\mu(\Delta t + o(\Delta t))$. In the fluid limit as n goes to infinity, the queue length and system time are scaled down by a factor n (see e.g., Jain *et al.* [2011]), the departure rate between time t and $t + \Delta t$ equals $\mu\Delta t + o(\Delta t)$, and the probability that a given customer in the queue departs in this interval is again $\frac{\mu}{c}\Delta t + o(\Delta t)$, as in (3).

It is also instructive to see that (3) may break down if the service requirement is not exponentially distributed. To see this simply, suppose that at time zero, n customers arrive to an empty system and thereafter there are no arrivals. Each

customer has service requirement of zero with probability $p \in (0,1)$ and $1/(1-p)$ with probability $(1-p)$ (so that the mean requirement is 1). Then, at time just after zero, the queue length reduces to $(1-p)n$. Each customer is served at rate $\frac{\mu}{(1-p)n}$ and hence all the remaining customers complete their service at time n/μ. In the corresponding limiting fluid model where queue length and system time are scaled down by n, one unit of customers arrive at time zero. Fraction p leave right away while the remaining $(1-p)$ leave at time $1/\mu$. Thus, average time spent in the system equals $(1-p)/\mu$. In particular, the customer waiting time is sensitive to the distribution of the service time and not just its mean. This makes the problem of analyzing processor sharing regime when the service times are generally distributed, even in the fluid setting, more challenging than in the exponential case that we consider.

Note that in this service regime, in equilibrium, customers have no incentive to come before time 0 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, in equilibrium, for:

- $\alpha \leq \beta$,
  - — All customers arrive at time 0.
- $\alpha > \beta$,
  - — At time zero, $\beta\Lambda/\alpha$ amount of customers arrive.
  - — The remaining $\Lambda(1-\beta/\alpha)$ customers arrive uniformly between times $[0, \Lambda/\mu]$.

We now illustrate this analysis.

3.1. $\alpha \leq \beta$

In this case, if all customers come at time zero then their average waiting times and times to service both equal $\Lambda/(2\mu)$ so their average cost is $(\alpha+\beta)\Lambda/(2\mu)$. 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 $0 < \delta \times \Lambda \leq \Lambda/\mu$. At time $\delta \times \Lambda$ there remain $\Lambda(1-\mu\delta)$ customers in queue. So this customer's average waiting time is

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

It will complete its service at an average time of

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

Therefore, its average cost is

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

which equals

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

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 socially optimal solution in this case remains the same as in FCFS, namely that customers arrive at rate μ during the interval $[0, \Lambda/\mu]$ so that there is no queue and the total social optimal cost equals $\beta\Lambda/(2\mu)$. It then follows that the price of anarchy equals $1/\beta$. This, surprisingly, is less than two when $\beta > \alpha$! This result was also observed in Haviv [2013] in the fluid models under the FCFS discipline when customers are not allowed to queue before time zero and $\alpha \leq \beta$. Hassin and Kleiner [2010] similarly note that restricting arrivals to not queue before time zero in an FCFS service regime reduces equilibrium cost compared to the unrestricted case (again under FCFS) for certain parameter values.

3.2. $\alpha > \beta$

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 Λ). It can be easily seen that again the server must serve at full rate till time Λ/μ, in particular $F(t) - \mu t > 0$ for $0 \leq t < \Lambda/\mu$ (else a customer can improve cost by coming at a time where the queue is empty). This also implies that $F(\Lambda/\mu) = 1$ so that $\mathcal{T} \subset [0, \Lambda/\mu]$. Further, the queue length $Q(t) = F(t) - \mu t$ for $0 \leq t < \Lambda/\mu$. In particular, the effective service rate for any customer present at time t, call it $\mu(t)$, for $0 < t < \Lambda/\mu$ 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 for any $t < \Lambda/\mu$, this does not put any mass beyond Λ/μ, we need to show that

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

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

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

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

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

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

$$EW_F(t) = \mu \int_t^{\Lambda/\mu} \frac{s - t}{F(s) - \mu s} \exp(-m_t(s)) ds. \tag{4}$$

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

The expected cost,

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

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

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

or

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

which is rearranged as

$$(\alpha + \beta) \left(\frac{\mu EW_F(t)}{F(t) - \mu t} - \int_t^{\Lambda/\mu} h_t(s) ds \right) + \beta = (\alpha + \beta) \left(\frac{\mu EW_F(t)}{F(t) - \mu t} - 1 \right) + \beta = 0. \tag{5}$$

Recall that $(\alpha+\beta) EW_F(t) = c_e - \beta t$ and substitute this in (5). Thus, equilibrium profile must satisfy

$$F(t) - \mu t = \mu \frac{c_e - \beta t}{\alpha} \tag{6}$$

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 $\alpha > \beta$ is necessary for $\mathcal{T}^o$ to be non-empty. It also follows that if $\alpha \leq \beta$ then $\mathcal{T}^o$ is empty. This is a key observation in establishing the uniqueness of the equilibrium profile proposed earlier for $\alpha \leq \beta$ (this needs to be coupled with the fact that a point mass arrival after time zero cannot be an equilibrium as a customer can then improve its cost by coming at time zero).

Number of observations can be seen to follow from (6):

(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 $0 \leq t_1 < t_2 < t_3 < t_4 \leq \Lambda/\mu$ 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 Juneja and Shimkin [2012] and other related references, we have limited ourselves to $\mathcal{T}$ that can be represented as a finite union of intervals along any bounded set.)

From (6) 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 (6) 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, suppose that $t_e = T < \Lambda/\mu$. Then, the customer that arrives at time T incurs a waiting time

$$EW_F(T) = \mu \int_T^{\Lambda/\mu} \frac{s - T}{\Lambda - \mu s} \exp(-m_T(s)) ds,$$

where

$$m_T(s) = \int_T^s \frac{\mu}{\Lambda - \mu y} dy = \log \frac{\Lambda - \mu T}{\Lambda - \mu s}.$$

Hence,

$$EW_F(T) = \frac{\mu}{\Lambda - \mu T} \int_T^{\Lambda/\mu} (s - T) ds = \frac{\Lambda - \mu T}{2\mu}.$$

In particular then

$$c_e = \frac{(\alpha+\beta)\Lambda}{2\mu} - \frac{(\alpha-\beta)T}{\mu}. \tag{7}$$

This is greater than the cost $\frac{\beta\Lambda}{\mu}$ incurred by a customer that arrives at time $\frac{\Lambda}{\mu}$, so that at equilibrium $t_e = T = \Lambda/\mu$.

Now we show that the profile necessitated by above discussion is indeed an equilibrium profile where each customer incurs an expected cost of $\beta\Lambda/\mu$.

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

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

for $0 \le t \le \Lambda/\mu$. $F(t) = 0$ for $t < 0$.

Then, for $0 \le t < \Lambda/\mu$,

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

This simplifies to equal

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

Then, for $0 \le t < \Lambda/\mu$

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

After some calculus, it can be seen that

$$EW_F(t) = (\Lambda/\mu - t)\frac{\beta}{\alpha+\beta}$$

so that $c_e = \beta\Lambda/\mu$. In particular, the proposed profile is indeed a unique equilibrium profile. Note that under the equilibrium profile, $F(0)$ equals $\beta\Lambda/\alpha$ so that this is the point mass of customers that arrive at time zero. As in the FCFS case, it is worth emphasizing that this arrival profile can be obtained in many different ways.

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

3.3. *Restricted arrival times*

Consider as in Haviv [2013] the case where the arrival times are restricted to come before some time T. Indeed if $T \le 0$, then the solution corresponds to everyone coming at time T as customers have no incentive to come earlier. When $T > 0$ and $\alpha \le \beta$, then as before, the equilibrium corresponds to all the customers arriving at time zero. When $\alpha > \beta$ and $T \ge \Lambda/\mu$ then under the equilibrium solution the arrival profile is as in the unrestricted case. When $T < \Lambda/\mu$, then carrying on our

analysis as before for $t \in (0, T)$, it follows that $F'(t) = (1-\beta/\alpha)\mu$. Since, $F(T) = \Lambda$, we get

$$F(t) = \Lambda - (T-t)(1-\beta/\alpha)\mu,$$

$F(t) = 0$ for $t < 0$ and $F(0) = \Lambda - T(1-\beta/\alpha)\mu$ denotes the point mass of arrivals at time 0. The equilibrium cost is then given by (7) and can be seen to be larger than $\beta\Lambda/\mu$.

The social welfare solution then depends on value of $T > 0$: If $T \geq \Lambda/\mu$, then as before arrivals come uniformly at rate μ between time 0 and Λ/μ. If, on the other hand, $T < \Lambda/\mu$, then arrivals come uniformly at rate μ between time 0 and T and then the remaining arrivals all come at time T.

4. Conclusion and Some Directions for Future Work

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 times and times to service want to join a queueing service that opens at a specified time and serves FCFS at a fixed rate.

Our chief contribution was to develop the equilibrium profile in the setting where the service discipline is no longer FCFS but is instead ROS or a limiting regime of processor sharing where the customer service times were independent and identically exponentially distributed. We found that in these cases, 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. We also discussed the arrival profile when the arrivals are restricted to come before a certain specified time.

Few issues immediately call for further research based on our analysis. We observed that our analysis is valid for processor sharing service regime, if the customer service times in the prelimit model are independent and identically exponentially distributed. We also observed that the situation maybe a great deal more complex when these service times are non-exponentially distributed. Analysis that arrives at the equilibrium profile under this general service times assumption would be practically quite useful. Furthermore, our analysis of ROS regime was greatly simplified because of the fluid model assumption. This may become much more complex when the population is finite and random. For instance, in the finite population FCFS case, equilibrium has to be symmetric (see Juneja and Shimkin [2012]). This profile is found by solving a functional differential equation. Finding socially optimal solution is also non-trivial problem in that setting. One expects similar difficulties in analyzing the finite case when service is in random order or if it follows processor sharing discipline. Thus, this generalization remains an interesting area for further research.

Acknowledgments

Authors would like to thank the editor and the anonymous referees for their comments that helped improve the manuscript.

References

Arnott, R., Palma, A. and Lindsey, R. [1999] Information and time-of-usage decisions in the bottleneck model with stochastic capacity and demand, *Eur. Econom. Rev.* **43**, 525–548.

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

Hassin, R. and Haviv, M. [2003] *To Queue or Not to Queue*, Equilibrium behavior in queneing systems, Vol. 59 (Springer, US).

Hassin, R. and Kleiner, Y. [2010] Equilibrium and optimal arrival patterns to a server with opening and closing times, *IIE Trans.* **43**(3), 164–175.

Haviv, M. [2013] When to arrive at a queue with tardiness costs? *Performance Evaluation* **70**(6), 387–399.

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

Jain, R., Juneja, S. and Shimkin, N. [2011] The concert queuing problem: To wait or to be late, *Discrete Event Dynam. Syst.* **21**, 103–138.

Juneja, S. and Jain, R. [2009] The concert/cafeteria queuing problem: A game of arrivals, *Proc. ValueTools'09 Fourth ICST/ACM Fourth Int. Conf. Performance Evaluation Methodologies and Tools*, Pisa, Italy.

Juneja, S. and Shimkin, N. [2012] The concert queueing game: Strategic arrivals with waiting and tardiness costs, *Queueing Syst.* **74**(4), 369–402, doi 10.1007/s11134-012-9329-3.

Juneja, S., Raheja, T. and Shimkin, N. [2012] The concert queuing game with random arrivals volume, *Proc. ValueTools'12 6th Int. Conf. Performance Evaluation Methodologies and Tools*, pp. 317–325.

Lindsey, R. [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.

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

Vickrey, W. S. [1969] Congestion theory and transport investment, *The American Economic Review* **59**, 251–260.

Wardrop, J. G. [1952] Some theoretical aspects of road traffic research, *Proc. Inst. Civil Engineers*, Part 2, Vol. 1, pp. 325–378.