Social Welfare and Price of Anarchy in Preemptive Priority Queues

Jonathan Chamberlain*, David Starobinski*

Department of Electrical and Computer Engineering, Boston University, 8 St Mary's Street, Boston MA 02215

Abstract

Consider an unobservable $M|G|1$ queue with preemptive-resume scheduling and two priority classes. Customers are strategic and may join the premium class for a fee. We analyze the resulting equilibrium outcomes, equilibrium stability, and social welfare. We find that for service distributions with coefficient of variation greater than 1, there exists a unique and stable mixed equilibrium at low loads. We also establish a tight bound on the price of anarchy, which is 4/3.

Keywords: Game Theory, Queuing Theory, Preemption, Pricing.

1. Introduction

Queuing models based on priority scheduling have been applied to a variety of contexts. These include transmission of multimedia traffic over a network [1], management of hospital beds and ambulances [2], and managing the smart grid [3]. All of these examples concern themselves with non-preemptive models. Yet, many other important applications, such as high performance computing [4] and scheduling of cloud containers [5], make use of *preemption*. As far as we are aware, however, there are relatively few studies of strategic customer behavior in preemptive queues.

While equilibria and social welfare within preemptive priority queues are considered in works such as [6, pp 83-85] [7] [8], the analyses either implicitly or explicitly assume that the $M|M|1$ regime is in effect. In particular, under a model in which customers may pay a fee to join the higher priority class such as in [6, pp 83-85], it is asserted that a mixed equilibrium state will never be stable, partly because if one exists it will not be the sole possible equilibrium state. In our paper, we show that this result does not extend to general service distributions.

*Corresponding author

Email addresses: jdchambo@bu.edu (Jonathan Chamberlain), staro@bu.edu (David Starobinski)

Specifically, we consider an $M|G|1$ queue with preemptive-resume scheduling discipline (i.e., preempted jobs resume from the point where they are interrupted), and two priority classes. Customers have the option on entering the queue to purchase access to a *premium* class, or otherwise remain in the *ordinary* class.

We use the standard notation λ to denote the mean arrival rate, μ for the mean service rate, and $\rho = \lambda/\mu$ for the traffic load. In addition, we use C to denote the cost to join the premium class, ϕ to denote the fraction of customers joining the premium class, and define a variance parameter K such that the second moment of service equals K/μ^2.

Under the model of an unobservable queue [6, pp 22, 53], we analyze the equilibrium outcomes, stability of the equilibria, and social welfare of the system. We show that the results are influenced both by the traffic load and the second moment of the service distribution. In particular, we show that a stable mixed equilibrium exists at sufficient low traffic load if $K > 2$ (i.e., the coefficient of variation is greater than 1). We further show that the price of anarchy of the queue is bounded by 4/3. These results stand in sharp contrast to the non-preemptive case where K bears no impact on the equilibrium outcomes and the price of anarchy is always 1.

2. Equilibrium Analysis

We are interested in the existence and stability of equilibria. In this model, the provider fixes the cost C to join the premium class prior to admitting customers. Thus, equilibria states are characterized by the fraction ϕ of customers who join the premium class. This results in three possible equilibria types:

1. *Everyone joins* the premium class, i.e. $\phi = 1$.
2. *No one joins* the premium class, i.e. $\phi = 0$.
3. *Some join* the premium class, i.e. $\phi \in (0, 1)$.

Equilibria of the first two types are *pure* strategies, as all customers make the same decision. Equilibria of the third type are *mixed* strategies, as a customer who is indifferent will join the premium class with probability ϕ and remain in the ordinary class with probability $1 - \phi$.

To show existence of possible equilibria, we must define where the customer will be indifferent between their options. As the customer chooses between the premium or ordinary class, they are indifferent if the costs of joining each class are identical. By assumption, customers are statistically identical, thus we need only consider the cost of waiting in the queue. By extension, WLOG we may assume a customer's cost of waiting in the queue equals the time spent waiting as the customers will have identical value on their time spent waiting. Letting $E[W_p]$ and $E[W_o]$ be the expected wait time in the queue as a member of the premium and ordinary classes respectively, a customer is indifferent if the following holds:

$$E[W_p] + C = E[W_o].$$

As the wait times depend on the fraction of customers in each class, we can relate a equilibrium strategy $\phi \in [0,1]$ to the cost which leads to that equilibrium by applying the formula for expected wait time in an $M|G|1$ priority-resume (PR) queue: [9, p.175]:

$$\mathcal{C}(\phi) \triangleq E[W_o] - E[W_p] = \frac{K\rho + (2-K)\phi\rho(1-\rho)}{2\mu(1-\rho)(1-\phi\rho)}. \tag{1}$$

We first evaluate the behavior of $\mathcal{C}(\phi)$ as ϕ increases from 0 to 1, to show how the relative costs of waiting changes as more customers attempt to join the premium class.

Lemma 1. *The function $\mathcal{C}(\phi)$, defined in Equation* (1)*, behaves as follows with respect to ϕ:*

1. *If $K > 2$ and $\rho < (K-2)/(2K-2)$, $\mathcal{C}(\phi)$ is monotone decreasing.*
2. *Else, if $K > 2$ and $\rho = (K-2)/(2K-2)$, $\mathcal{C}(\phi)$ is constant valued.*
3. *Otherwise, $\mathcal{C}(\phi)$ is monotone increasing.*

The proof is obtained by computing the derivative $\mathcal{C}'(\phi)$ and determining the conditions for which it is positive, negative, or zero. It turns out that the sign of the derivative is determined by the sign of its numerator, which is constant with respect to ϕ. Therefore, $\mathcal{C}(\phi)$ must be monotone or constant, the exact behavior depending on the values of K and ρ, and the rest follows.

As $\mathcal{C}(\phi)$ must be monotone or constant, $\min \mathcal{C}(\phi)$ and $\max \mathcal{C}(\phi)$ will be well defined quantities. Further, each of these will be equal to either $\mathcal{C}(0)$ or $\mathcal{C}(1)$ depending on the exact behavior of $\mathcal{C}(\phi)$. If $\mathcal{C}(\phi)$ is monotone, there is a unique solution to $\mathcal{C}(\phi) = C$, which is

$$\phi_e = \frac{2\mu C(1-\rho) - K\rho}{\rho(1-\rho)(2\mu C + 2 - K)}. \tag{2}$$

Thus, given a cost C, we determine the existence of possible equilibria by relating C to $\mathcal{C}(\phi)$. When evaluating the stability of any such possible equilibria, we apply the Evolutionary Stable Strategy definition from [10]:

Definition 1. *A strategy adopted by a population which cannot be invaded by an initially rare strategy is said to be an* Evolutionary Stable Strategy (ESS)*. That is, if strategy ϕ^* is ESS, then no other equilibrium strategy ϕ^{**} exists such that ϕ^{**} is a best response to ϕ^*.*

We next present the main results of this section:

Theorem 1. *The equilibria of a two-class $M|G|1$-PR queue have the following structure:*

1. *If $C < \min \mathcal{C}(\phi)$,* everyone joins *is the unique equilibrium.*
2. *If $C > \max \mathcal{C}(\phi)$,* no one joins *is the unique equilibrium.*

3. *If* $\min \mathcal{C}(\phi) < C < \max \mathcal{C}(\phi)$ *and* $\mathcal{C}(\phi)$ *is monotone decreasing, a* some join *equilibirum with* ϕ_e *customers in the premium class is the unique equilibrium.*
4. *If* $\min \mathcal{C}(\phi) < C < \max \mathcal{C}(\phi)$ *and* $\mathcal{C}(\phi)$ *is monotone increasing, the* everyone joins *equilibrium,* no one joins *equilibirum, and* some join *equilibrium with* ϕ_e *customers are all possible equilibria.*

Pure equilbibria are always ESS. The mixed equilibrium is ESS if and only if it is unique.

Corollary 1. *By Lemma 1 and Theorem 1, a unique mixed equilibrium exists if and only if* $K > 2$ *and* $\rho < (K-2)/(2K-2)$*. Furthermore, this equlibrium is ESS.*

We next sketch the proof of Theorem 1. The first two cases follow trivially from the fact that if C is smaller than the minimum (or respectively, greater than the maximum) joining the premium class will cost less (resp., more) than remaining in the ordinary class regardless of how many customers are in the premium class. Therefore, the only possible equilibria is the *everyone joins* (resp., *no one joins*) equilibrium.

In the third case, ϕ_e being a possible equilibrium follows by definition. That no others are possible follows from $\mathcal{C}(\phi)$ being monotone decreasing: if more customers join the premium class than under equilibrium, then it is cheaper to remain in the ordinary class, and vice versa. Hence, a deviation from ϕ_e has a best response of pushing the system back to ϕ_e, and no other equilibrium is possible.

In the fourth case, ϕ_e is again a possible equilibrium state by definition. However, because $\mathcal{C}(\phi)$ is monotone increasing, if more customers join the premium class than under equilibrium, then it is cheaper to join the premium class. The reverse is also true if more customers join the ordinary class than under equilibrium. This results in the system reaching a pure state if deviating from ϕ_e. These pure states are also possible equilibria as once in a pure state, it costs more to attempt to join the opposite class of what all other customers have chosen.

The ESS criteria follows as a corollary from the previous assertions. In the first three cases, deviating from the equilibrium is never a best response. In the fourth case, pure equilibria are possible by deviating from the mixed equilibrium.

3. Social Welfare Analysis

We now shift our analysis to the social welfare and attendant price of anarchy [11]. Social welfare is defined in terms of the utilities of the customers and the provider. However, here customers are statistically identical, and the preemption policy is work-conserving. Further, the cost C to join the premium class is a transfer from customers to the provider. As a result, the social welfare only varies based on the costs of waiting in the queue for service. Thus we maximize the social benefit by minimizing overall wait times. We derive the

expected average wait time $E[W]$ from the expected wait times in each priority class as follows:

$$E[W] = \frac{\rho\left(K - 2\phi\rho + (2-K)\phi(1-\phi(1-\rho))\right)}{2\mu(1-\rho)(1-\phi\rho)}. \tag{3}$$

Lemma 2. *Let ϕ^* be defined as follows:*

$$\phi^* \triangleq \frac{1-\sqrt{1-\rho}}{\rho}. \tag{4}$$

In this model, the socially optimal states depend on the value of K as follows:

1. *If $K < 2$, the socially optimal states are $\phi = 0$ and $\phi = 1$;*
2. *If $K = 2$, all states $\phi \in [0, 1]$ are socially optimal;*
3. *If $K > 2$, the socially optimal state is $\phi = \phi^*$.*

To prove this, we compute the derivative of $E[W]$ with respect to ϕ. In doing so, we find that the sign of the derivative flips when $\phi = \phi^*$. If $K < 2$, the sign of the derivative flips from positive to negative at ϕ^*, thus ϕ^* results in the maximum possible wait time, and conversely $\phi = 0$ and $\phi = 1$ result in the minimum wait time. If $K = 2$, then the derivative is 0 for all ϕ, thus the wait time is constant with respect to ϕ and all states are optimal by default. If $K > 2$, the sign of the derivative flips from negative to positive at ϕ^* and therefore ϕ^* results in the minimum possible wait time.

As K is defined in terms of the second moment, this means that if variance in service is less than that of exponential, the social welfare is maximized when all customers join the same class. If variance is greater than that of exponential, the social welfare is maximized when a specific fraction ϕ^* of customers join the premium class. If the service distribution is exponential however, then it does not matter how many customers join the premium class, as the choice to join or not does not impact the overall welfare. This can ultimately be shown to result from the relation between the expected time of service of a preempting customer $(1/\mu)$ and the expected residual time of service of the preempted customer $(K/(2\mu))$.

3.1. Price of Anarchy

The Price of Anarchy (PoA) is a measure of the loss of optimality resulting from a lack of cooperation. This is defined by comparing the socially optimal state to the equilibrium state which leads to the highest social cost [11]. As the social welfare depends solely on $E[W]$, we define the PoA for our system in terms of the costs of waiting:

Definition 2. *Let $E \subset [0, 1]$ be the set of possible equilibria for fixed cost C and traffic load ρ. The Price of Anarchy (PoA) is defined as the following ratio:*

$$PoA = \frac{\max\limits_{\phi \in E} E[W]}{\min\limits_{\phi \in [0,1]} E[W]}.$$

As possible equilibria and the socially optimal state depend on K, ρ, and C, so must the PoA. As a result, we aim to show that an upper bound exists on the PoA, such that there is a clearly defined *worst case* scenario which cannot be exceeded.

Theorem 2. *The price of anarchy of a two-class $M|G|1$-PR queue is bounded from above by* $4/3$.

To prove that such an upper bound exists, we assume that given arbitrary but fixed ρ and K, C is set such that the state ϕ which leads to the largest possible average wait time $E[W]$ is in the set of possible equilibria. Per Lemma 2, we note that if $K < 2$, the socially optimal states are $\phi \in \{0, 1\}$. Using the proof of the lemma, it is straightforward to show that ϕ^* is the corresponding worst-case state. This results in

$$PoA = \frac{(2-K)\left(2 - 2(1-\rho)^{\frac{3}{2}} - 3\rho\right)}{K\rho^2} + \frac{2}{K}. \tag{5}$$

If $K = 2$, then all states are socially optimal, and thus so are all possible equilibria states. It then follows that if $K = 2$, the PoA is equal to 1, regardless of the value of ρ.

If $K > 2$, then the socially optimal state is $\phi = \phi^*$, and the worst case state is $\phi \in \{0, 1\}$. Thus, we have a PoA which is the reciprocal of expression in Equation (5):

$$PoA = \frac{K\rho^2}{(2-K)(2 - 2(1-\rho)^{\frac{3}{2}} - 3\rho) + 2\rho^2}. \tag{6}$$

Thus, to determine an upper bound on the price of anarchy, we determine the suprema of Equations (5) and (6) given $\rho \in (0, 1)$, and the respective bounds on K for each equation. We observe that Equation (5) will be maximized when $K = 1$; evaluating with respect to ρ we find that the PoA when $K < 2$ is bounded from above by $5/4$, as ρ approaches 0. Evaluating Equation (6), we find that it too will be maximized as ρ approaches 0, resulting in a bound of

$$PoA = \frac{4K}{2 + 3K}.$$

And as K approaches infinity, this quantity is bounded from above by $4/3$. Thus, regardless of the values of K, C, or ρ, the price of anarchy is never greater than $4/3$. This bound is reached as ρ approaches 0. Thus, low traffic loads lead to the greatest cost from a lack of cooperation. Conversely, as $\rho \to 1$, the PoA approaches 1.

4. Comparison to the Non-Preemptive (NP) Queue

We now briefly contrast the results to the situation where customers are offered the ability to purchase access to the premium class, but no customers

may be preempted while in service. Based upon the formula for waiting in an $M|G|1$ queue with no preemption [9, p.164], one can derive the resulting cost function $\mathcal{C}(\phi)$ and expected average wait time $E[W]$ as follows:

$$\mathcal{C}(\phi) = \frac{K\rho^2}{2\mu(1-\rho)(1-\phi\rho)},$$
$$E[W] = \frac{K\rho}{2\mu(1-\rho)}.$$

We immediately note that $E[W]$ is constant with respect to ϕ, thus the average wait time does not depend on the number of customers in each class. This results in all states being socially optimal and the $PoA = 1$ by default, regardless of which equilibria are possible. Evaluating $\mathcal{C}(\phi)$, we find that in a non-preemptive queue, the following equilibria structure prevails:

- If $C < \mathcal{C}(0)$, *everyone joins* is the only possible equilibrium.
- If $C > \mathcal{C}(1)$, *no one joins* is the only possible equilibrium.
- If $\mathcal{C}(0) < C < \mathcal{C}(1)$, there are three possible equilibria: *everyone joins*, *no one joins*, and a *some join* with $\phi = 1/\rho - (K\rho)/(2\mu C(1-\rho))$.

The pure equilibria are ESS, while the *some join* will never be ESS. Thus, in contrast to the preemptive queue, customers will tend towards all joining the same class no matter what the service variance or traffic load are. Indeed, the service variance will only impact the cost that can be charged to join the premium class, but it has no impact on the social welfare.

5. Conclusions

In this paper we analyzed the equilibrium outcomes of a two-class $M|G|1$-PR queue where customers can purchase access to the higher priority class for a fee. We find that if the variance in service is greater than that of the exponential distribution and the traffic load is sufficiently low, then there exist conditions under which a stable mixed equilibrium is possible. This is not possible otherwise (e.g., for exponential and deterministic service times).

We also conducted a social welfare analysis and showed that a mixed state is socially optimal if service variance is sufficiently high (i.e., $K > 2$). Thus, it is not the case generally that the system is always best off when all customers join one class or the other, as might be observed when looking at more deterministic systems. In any event, the resulting price of anarchy is bounded by 4/3. As the traffic load increases to 1, the price of anarchy tends to 1. Finally, we observe that the social welfare in the $M/G/1$-PR queue can either be smaller or larger that of the $M/G/1$-NP queue. However, the former case is likelier because only in the $M/G/1$-PR queue a mixed equilibrium is stable, and this can only be reached when $K > 2$ (i.e., when a mixed equilibrium leads to a higher social welfare than the pure equilibria).

Acknowledgements

This work was partially supported by NSF grants 1717858 and 1908087.

Bibliography

[1] J. Walraevens, B. Steyaert, H. Bruneel, Performance Analysis of the System Contents in a Discrete-Time Non-Preemptive Priority Queue With General Service Times, Belgian Journal of Operations Research, Statistics and Computer Science (JORBEL) 40 (2000) 91–103.

[2] E. A. Peköz, Optimal Policies for Multi-Server Non-Preemptive Priority Queues, Queueing Systems 42 (1) (2002) 91–101.

[3] S. Sadeghi, M. H. Yaghmaee Moghddam, M. Bahekmat, A. S. Heydari Yazdi, Modeling of Smart Grid Traffics Using Non-Preemptive Priority Queues, in: 2012 2nd Iranian Conference on Smart Grids, ICSG 2012, IEEE, 2012, pp. 1–4.

[4] L. Jin, Cook: A Fair Preemptive Resource Scheduler for Compute Clusters (2016) [cited 2019-04-05].
URL `https://www.twosigma.com/insights/article/cook-a-fair-preemptive-resource-scheduler-for-compute-clusters/`

[5] R. Hat, Pod Priority and Preemption - Scheduling | Cluster Administration | OpenShift Container Platform 3.11 (2020) [cited 2020-02-03].
URL `https://docs.openshift.com/container-platform/3.11/admin_guide/scheduling/priority_preemption.html`

[6] R. Hassin, M. Haviv, To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems, Kulwer Academic, Boston, 2003.

[7] Y. Shi, C.-f. Wang, Analyses of Equilibrium Behaviour of Customers and Optimal Design for Queues Under a Preemptive Priority Discipline, in: Proceedings of the 2019 3rd International Conference on Management Engineering, Software Engineering and Service Sciences, ICMSS 2019, ACM, 2019, pp. 93–96.

[8] C. O. Itai Gurvich, Martin A. Lariviere, Coverage, Coarseness, and Classification: Determinants of Social Efficiency in Priority Queues, Management Science 65 (3) (2019) 1061–1075.

[9] R. Conway, W. Maxwell, L. Miller, Theory of Scheduling, Addison-Wesley, Reading, MA, 1967.

[10] J. M. Smith, Evolution and the Theory of Games, Cambridge University Press, 1982.

[11] G. Gilboa-Freedman, R. Hassin, Y. Kerner, The Price of Anarchy in the Markovian Single Server Queue, IEEE Transactions on Automatic Control 59 (2) (2014) 455–459.