Pricing for congestion in an M/G/1 queue [1]

Moshe Haviv[2] [3] [4] and Ya'acov Ritov[3] [5]

Abstract

No doubt that a customer (or a job) who joins a queue imposes delays on others. He should be charged for that. The question then is by how much. In particular, one looks for a price mechanism with which total waiting costs is shared while customers pay in accordance with their own length of service. Two such price mechanisms are presented here. The first, based on the notion of *externalities*, is by charging customers for the total waiting time accumulated while they are being served. The second is based on applying the *Aumann-Shapley cost-sharing price mechanism*. These two mechanisms will be exemplify in the M/G/1 queueing model with various entrance disciplies like first-come first-served and processor sharing.

1 Introduction

Many cases in which users share a common facility can be modeled as queueing systems. This is, for example, the case when airplanes take off from a

[1] This paper summarizes some of the content of two papers. The first, titled *Externalities, tangible externalities and queue disciplies*, written by the authors, appeared in *Management Science*, Vol. 44, pp. 850–858, 1998 and it content is discussed in Section 2. The second, *The Aumann-Shapley price mechanism for allocating costs in congested systems*, written by the first author, was submitted for publication and its content is dicussed in Section 3. No proofs are supplied here and they can be found in the abovementioned papers.

[2] Department of Econometrics, The University of Sydney, Sydney NSW 2006, Australia.

[3] Department of Statistics, The Hebrew University of Jerusalem, 91905 Jerusalem, Israel.

[4] E-mail: mosheh@econ.usyd.edu.au.

[5] E-mail: yaacov@mscc.huji.ac.il.

runway, when commuters cross a bridge, when jobs share a common CPU and when messages are routed through the same data network. In such systems users are being delayed due to the presence of others and likewise users themselves usually impose delays on others. These delays can be looked at as *congestion costs* which are imposed on (or by) individual users on (or by) the entire population of users. See [5] for more on this notion. The question we looked at here is how users, or types of users, should be charged for their fair share in the congestion costs that they impose on the society.

The amount of externalities that a user imposes on others is typically a function of his own requirement. Specifically, in the context of service systems, the longer his service requirement is (or its expected value, resp.), the larger are the negative externalities (or their expected value, resp.) he imposes. Defining quantitatively the externalities can serve for our main purpose here which is to suggest a queue manager how much to charge users for the use of the common facility. In such schemes is makes sense to do so in accordance with the congestion costs that an individual imposes on others. This is of course on top of other charges such as access charges, user charges or service quality charges. See [5] for details.

Using this approach we suggest is Section 2 two possible mechanisms. One is called *expected externalities* and the other is called *expected tangible externalities*. We consider them in the case of an M/G/1 queue and they are looked as functions of actual service requirements. We like to point out that the second mechanism is a cost-sharing procedure while the first one is not (actually the total charge it implies exceeds the total cost). Exact functions are derived under various queue disciplines like first-come first-served and processor sharing.

In Section 3 another approach is suggested. Again, the M/G/1 queueing model is used. There each customer is assumed to belong to one out of finitely many classes. Classes may vary with respect to the arrival rate and distribution of service requirement. Here we look at the costs that a class inflicts on the entire society. We then show how the Shapley-Aumann price mechanism can be used in order to measure these costs. The resulting prices have the cost sharing property. The mechanism also yields prices to charge individual with when one assumes that the arrival rate of a class degenerates to zero. As for the first price mechanism, formulas for the prices are developed for M/G/1 queues under various queueing disciplines like those mentioned above.

2 Expected externalities and tangible externalities

2.1 Externalities and tangible externalities

The standard definition for externalities that an individual imposes on others is the difference between the actual total queueing time of the others and the corresponding value in a (simulated) realization in which everything stays the same but without the presence of the user in question. By saying that everything stays the same we mean that all other arrivals occur at the same instants and with the same workload requirements.

Externalities as defined above do not possess the cost-sharing property, namely, their sum does not coincide with the total queueing time. Actually, this sum overestimates the total queueing time. For example, consider a single server first come first served (FCFS) queue where three customers A, B and C with identical service requirements are the first to arrive, and in this order, at an empty queue. Also, assume that the arrival times of customers A and B (almost) coincide, while customer C arrives when customer A is ready to clear the system. Hence, the externalities of *both* customers A and B include the waiting time of customer C.

We suggest here an alternative measure. For each customer we associate what we call *tangible externalities*. They are defined as the total queueing time added to the others *while this customer is in service*. This measure possesses the cost-sharing property since any part of the total queueing time is now associated with one and only one customer. In particular, the sum of tangible externalities across customers equals the total queueing time.

The above suggested mechanisms, used as defined, may seem a bit unfair for some customers. For example, in a FCFS system, a customer who is asked to pay a bundle may claim that it was not his fault that many customers arrived during his service period: He was served during a particular period in which the number of arrivals was high by a mere chance or just because he was delayed service by other customers who have arrived earlier. Of course, other customers may be more lucky. Better measures respectively are *expected externalities* and the *expected tangible externalities* which a customer imposes as functions of his (actual) service requirement. These measures have the same uncondional expected value (over all customers) as their counterparts,

but their variances are respectively smaller.

Besides variance reduction, there is an additional advantage to use expected values. This is the simplicity of using them as there is no need to monitor how many customers are in the system upon each service commencement and how many arrive and at exactly which instants during this service period. All that is required is to monitor the length of the service requirements themselves. Actually, the queue operator can place a sign which tells each customer how much to pay as a function of his service requirement.

Before moving on we need the following definitions and notation. Recall that M/M/1 stands for a single server queue with a Poisson arrival process and exponential service requirement. The waiting room is unlimited and all random variables involved are independent and the system is considered to be in steady state. The entrance to service discipline is not specified yet. Let λ be the rate of arrival. Each arrival requires an exponential amount of service with an expected value of μ^{-1}. Denote λ/μ by ρ, which is referred to as the *traffic intensity*. Of course, $0 \leq \rho < 1$. Finally, M/G/1 queues are defined as M/M/1 queues but without the assumption of exponential service time. In this case too ρ stands for the traffic intensity, namely $\lambda\kappa_1$ where κ_1 is the expected service requirement.

Call a customer whose service requirement is x an *x-customer* and let $T|x$ be the (random) externalities that he imposes on others. Then, let $\mathrm{E}(T|x)$ be the corresponding conditional expected value and let $\mathrm{Var}(\mathrm{E}(T|x))$ be the variance of the conditional expected values. Note that $\mathrm{Var}(\mathrm{E}(T|x))$ (as oppose to $\mathrm{Var}(T)$), for example) is the right measure for the variability for the prices paid by customers in the case that the price mechanism of externalities is adopted. Finally, let $C|x$ be the (random) tangible externalities that he imposes on others. Then, all of the above is defined in a similar way with respect to this quantity when C replaces T.

2.2 Expected externalities

2.2.1 The case of $M/G/1$ queues without preemption

We need the following notation. Denote by κ_n the n-th moment of service requirement. In particular ρ equals $\lambda\kappa_1 < 1$. Note that in the case of exponential service $\kappa_n = n!/\mu^n$.

Theorem 2.1 *For any strong and work-conserving M/G/1 queue without preemption,* [6] *in particular for first-come first-served (FCFS), non-preemptive last-come first-served (LCFS) and non-preemptive random M/G/1 queues,*

$$\mathrm{E}(T|x) = \frac{\lambda}{2(1-\rho)}x^2 + \frac{\lambda^2\kappa_2}{2(1-\rho)^2}x \tag{1}$$

and

$$\mathrm{Var}(\mathrm{E}(T|x)) = \frac{\lambda^2}{4(1-\rho)^2}(\kappa_4 - \kappa_2^2) + \frac{\lambda^3\kappa_2}{2(1-\rho)^3}(\kappa_3 - \kappa_2\kappa_1) + \frac{\lambda^4\kappa_2^2}{4(1-\rho)^4}(\kappa_2 - \kappa_1^2) \quad .$$

2.2.2 The LCFS-PR M/G/1 queue and PS M/M/1 queue

Here we consider two service disciplines. The first is the last-come first-served preemption-resume (LCFS-PR) discipline (also called a *stack*). [7] The second is the processor-sharing (PS) queueing model. [8]

---

[6] A queueing discipline is said to be *strong* if the order in which customers receive service is not a function of their actual service requirements (i.e., future values from the operator's point of view). Put differently, the queue operator is not informed of these values while deciding on who gets service when. Also, a queueing system is said to be *work-conserving* if the resulting process of total (residual) service requirement due to customers currently in the system (also known as the *virtual waiting time*) coincides with that of a FCFS system having the same characteristics.

[7] Here an arrival starts receiving service as soon as he arrives, possibly preempting the customer who is in service. Preempted customers return to service in a reverse order of their arrival. Finally, service is resumed from the point where it was interrupted.

[8] Here the server, at any time, shares evenly its capacity among all customers present in the system. This model is the limit of the round-robin scheme. Specifically, in a round-robin scheme all customers present in the system alternate receiving service of some quantum. So if the time length of a cycle is Δ and a fixed number of $n+1$ customers are present during this cycle, then each of them receives an amount of service of $\Delta/(n+1)$ (and imposes tangible externalities of $n\Delta/(n+1)$ on the others). Note that in order to receive Δ units of service, a customer has to spend $(n+1)\Delta$ units of time in the system. The processor sharing model corresponds to the case where Δ goes to zero. A way to interpret this model is by assuming that a customer has his own watch which runs only while he is in service. In particular, an x-customer leaves when his watch says x. In comparison with the natural clock, this watch progresses at a variable pace: when n additional customers are in the system, his watch moves $(n+1)$ times slower.

Theorem 2.2 *For the LCFS-PR M/G/1 queueing system and for the PS M/M/1 queueing system*

$$\mathrm{E}(T|x) = \frac{\rho}{(1-\rho)^2}x$$

and

$$\mathrm{Var}(\mathrm{E}(T|x)) = \frac{\rho^2}{(1-\rho)^4}(\kappa_2 - \kappa_1^2) \quad .$$

2.3 Expected tangible externalities

Here we state, $\mathrm{E}(C|x)$, the expected tangible externalities incurred by an x-customer and the corresponding variance $\mathrm{Var}(\mathrm{E}(C|x))$. We like to note that all expected values are smaller by a factor of $1-\rho$ of their counterparts in the previous subsection. However, the conjecture that for any queue discipline $\mathrm{E}(C|x) = (1-\rho)\,\mathrm{E}(T|x)$ is false.

2.3.1 M/G/1 queues without preemption

Theorem 2.3 *For any strong work-conserving M/G/1 queue without preemption, in particular for FCFS, LCFS and random order $M/G/1$ queueing systems,*

$$\mathrm{E}(C|x) = \frac{\lambda^2\kappa_2}{2(1-\rho)}x + \frac{\lambda}{2}x^2 \tag{2}$$

and

$$\mathrm{Var}(\mathrm{E}(C|x)) = \frac{\lambda^4\kappa_2^2}{4(1-\rho)^2}(\kappa_2 - \kappa_1^2) + \frac{\lambda^2}{4}(\kappa_4 - \kappa_2^2) + \frac{\lambda^3\kappa_2}{2(1-\rho)}(\kappa_3 - \kappa_2\kappa_1) \tag{3}$$

2.3.2 LCFS-PR and PS M/G/1 queues

Theorem 2.4 *For a LCFS-PR and PS M/G/1 queues,*

$$\mathrm{E}(C|x) = \frac{\rho}{1-\rho}x \tag{4}$$

and

$$\mathrm{Var}(\mathrm{E}(C|x)) = \frac{\rho^2}{(1-\rho)^2}(\kappa_2 - \kappa_1^2) \quad . \tag{5}$$

3 The Aumann-Shapley price mechanism

3.1 Definition

We next introduce notation to be used throughout this section. There are n types of customers with the following properties:

1. The arrival process of type-i customers is Poisson with rate λ_i, $1 \leq i \leq n$. Let $\lambda = \Sigma_{i=1}^{n} \lambda_i$ and let $\underline{\lambda} \in R^n$ be the vector of arrival rates, i.e., $\underline{\lambda} = (\lambda_1, \ldots, \lambda_n)$;

2. Denote the first and the second moments of the service requirement of type-i customers by κ_i and $\kappa_i^{(2)}$, respectively, $1 \leq i \leq n$;

3. Let $\rho_i = \lambda_i \kappa_i$, $1 \leq i \leq n$, let $\rho = \Sigma_{i=1}^{n} \rho_i$ and finally let $\rho^{(2)} = \Sigma_{i=1}^{n} \lambda_i \kappa_i^{(2)}$.

Consider an M/G/1 queueing system with n types of customers. The queue discipline is not specified yet: Various queue disciplines will be dealt with later. Let $W_j(\underline{\lambda})$, $j = 1, \ldots, n$, be the expected queueing time (excluding service) of a type-j customer when $\underline{\lambda}$ is the vector of arrival rates.

The total waiting cots per unit of time is

$$W(\underline{\lambda}) = \Sigma_{i=1}^{n} \lambda_i W_i(\underline{\lambda}) \tag{6}$$

and here, as in the previous section, we look for a cost sharing mechanism for splitting (6) among the various types of customers.

The Aumann-Shapley price mechanism says that

$$P_i(\underline{\lambda}) \equiv \int_{t=0}^{1} \frac{\partial W}{\partial \lambda_i}(t\underline{\lambda}) d\, t \tag{7}$$

is type-i's share of $W(\underline{\lambda})$ per unit of its arrival rate. Note (7) is a simple averaging of the directional derivative of the variable in question, hence it measures its average 'contribution' in the growth of the function $W(.)$ from $W(\underline{0})$ to $W(\lambda)$. Likewise,

$$\lambda_i P_i(\underline{\lambda}) \tag{8}$$

is the share in (7) of type-i. In other words, (8) is what type-i has to pay per unit of time and by the cost-sharing property

$$W(\underline{\lambda}) = \sum_{i=1}^{n} \lambda_i P_i(\underline{\lambda}) \quad .$$

Finally, note that $P_i(\underline{\lambda})$ is measured by the same units as time is measured with and that the price mechanism suggested in (7) implicitly charges a type also for (some of) the waiting time of its own members.

The A-S price mechanism suggested here can be extended and use for the charging of individual customers whom service requirements are known. This, of course, leads to an alternative scheme to the scheme of tangible externalities defined in the previous section. Specifically, suppose to the system described above a customer whose service requirement is x, is added. The price suggested here is

$$P_x(\underline{\lambda}) \equiv P_{n+1}(\underline{\lambda}, 0)$$

where $\kappa_{n+1} = x$ and $\kappa_{n+1}^{(2)} = x^2$. Special interest exists when $n = 1$ and $\underline{\lambda} = \lambda$. In this case the selected individual can actually be any one of this single type of customers!

The Aumann-Shapley price mechanism shares many other appealing properties (on top of being a cost-share procedure). Moreover, under a suitable selection of the domain of cost functions, this mechanism is sometimes shown to be the unique one which obeys some subset of these properties (called *axioms* in this context). See [1] or [4] for more details.

3.2 The prices

3.2.1 M/G/1 queues without preemption

The famous Pollaczek-Khinchin formula, we get that the expected waiting time regardless of type is

$$W(\underline{\lambda}) = \lambda \frac{\rho^{(2)}}{2(1-\rho)}$$

from which the next theorem follows.

Theorem 3.1 For non-preemptive M/G/1 queues, like FCFS, LCFS and random queues,

$$P_i(\underline{\lambda}) = -\frac{\rho^{(2)} + \lambda\kappa_i^{(2)}}{2\rho^2}(\rho + \log(1-\rho)) - \frac{\lambda\rho^{(2)}\kappa_i}{2\rho^3}(1 - \rho - \frac{1}{1-\rho} - 2\log(1-\rho)) \ , \ 1 \le i \le n \ . \tag{9}$$

Corollary 3.1 For the above queues,

$$P_x(\underline{\lambda}) = -\frac{\rho^{(2)} + \lambda x^2}{2\rho^2}(\rho + \log(1-\rho)) - \frac{\lambda \rho^{(2)} x}{2\rho^3}(1 - \rho - \frac{1}{1-\rho} - 2\log(1-\rho)) \quad (10)$$

Note that in this corollary $\underline{\lambda}$ can be replaced by λ as $P_x(\underline{\lambda})$ is a function of $\underline{\lambda}$ only through λ. Note that the prices given in (10) are an efficient way for sharing the waiting costs among customers, where those who require longer service pay more. It is easy to see that $P_x(\underline{\lambda})$ is a quadratic function in x, say $a_2 x^2 + a_1 x^1 + a_0$, with all three coefficients a_0, a_1 and a_2 being positive. In particular, $P_0(\underline{\lambda})$ is positive as it equals $-\frac{\rho^{(2)}}{2\rho}(\rho + \log(1-\rho))$. Maybe at first sight one likes to charge a zero customer (i.e., a customer who requires a service of length zero) by zero. However, in a second thought charging a zero customer should not be ruled out: even customers who require no service at all, just by their arrival add congestion and their waiting time is added to the total social waiting time. Note, however that in accordance with the price mechanism defined in the previous section, (see (2)) a zero customer, and likewise a type consisting of zero customers, pay zero.

3.2.2 The LCFS-PR and PS M/G/1 queues

For the LCFS-PR and PS M/G/1 queueing systems (see [3]),

$$W_i(\underline{\lambda}) = \frac{\rho}{1-\rho}\kappa_i \quad , \quad 1 \le i \le n$$

and hence one concludes that

$$W(\underline{\lambda}) = \frac{\rho^2}{1-\rho} \quad .$$

Theorem 3.2 In the case of a LCFS-PR and PS M/G/1 queues,

$$P_i(\underline{\lambda}) = \frac{\rho}{1-\rho}\kappa_i \quad , \quad 1 \le i \le n \quad (11)$$

Corollary 3.2 In the above-mentioned queues,

$$P_x(\underline{\lambda}) = \frac{\rho}{1-\rho}x$$

References

[1] Aumann, R.J. and L.S. Shapley (1974), *Values of Non-Atomic Games*, Princeton University Press, Princeton, New Jersey.

[2] Kleinrock, S. (1976), *Queueing Systems, Vol 2.: Computer Applications*, Wiley-Interscience, New York.

[3] Ross, S.M. (1983), *Stochastic Processes*, John Wiley & Sons, New York.

[4] Tauman, Y. (1988), "The Aumann-Shapley prices: a servey," in *The Shapley Value*, edited by A. Roth, Cambridge University Press.

[5] Walrand, J. and P. Varaiya (1996), *High–Performance Communication Networks*, Morgan Kaufmann Publishers, San Francisco.