Available online at www.sciencedirect.com
ScienceDirect
European Journal of Operational Research 188 (2008) 485–495
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/ejor

Stochastics and Statistics

Cooperation in Markovian queueing models ☆

M.D. García-Sanz [a], F.R. Fernández [b], M.G. Fiestras-Janeiro [c,*], I. García-Jurado [d], J. Puerto [b]

[a] *Faculty of Economics and Business, Salamanca University, 37007 Salamanca, Spain*
[b] *Department of Statistics and OR, Faculty of Mathematics, Sevilla University, 41012 Sevilla, Spain*
[c] *Department of Statistics and OR, Faculty of Economics, Vigo University, 36271 Vigo, Spain*
[d] *Department of Statistics and OR, Faculty of Mathematics, Santiago de Compostela University, 15782 Santiago de Compostela, Spain*

Received 17 February 2006; accepted 20 April 2007
Available online 25 May 2007

Abstract

In this paper we study some cooperative models in Markovian queues. We stress the case of several agents agreeing to maintain a common server for their populations in which a priority scheme with preemption has been established. In this situation we propose and characterize an allocation rule for the holding costs that provides core allocations.

Keywords: Markovian queues; Cooperative games; Allocation rules

1. Introduction

The study of cooperation and competition in operational research models is a fruitful and challenging topic nowadays. Most fields within operations research are being approached from a game theoretical perspective, for the cases in which several decision makers interact in situations that can be modeled as optimization problems. Borm et al. (2001) provides a review of this topic.

One of the major branches within operations research is *queueing theory*. Competition in queueing models has been treated in many papers, a survey of which is Hassin and Haviv (2003) (for a survey in the control of queues, the reader is referred to Tadj and Choudhury (2005)). There are also a number of papers on cooperative issues in sequencing and scheduling (see, for instance, a review in Curiel et al. (2002) or other recent references such as Moulin and Stong (2002) and Maniquet (2003)). However, surprisingly enough, queueing

☆ The authors acknowledge the financial support of European Science Foundation, *Ministerio de Educación y Ciencia*, FEDER, *Junta de Andalucía*, *Xunta de Galicia*, and *Junta de Castilla-León* through projects SEC2002-10181-E, MTM2004-0909, SEJ2005-07637-C02-02, SEJ2005-0304/ECON, HA2003-0121, FQM331-JuntaAndalucía, PGIDIT06PXIC207038PN, and SA098A05. The helpful comments of three anonymous referees are also acknowledged.
* Corresponding author. Tel.: +34 986812498.
E-mail address: fiestras@uvigo.es (M.G. Fiestras-Janeiro).

0377-2217/$ - see front matter
doi:10.1016/j.ejor.2007.04.053

models have rarely been approached from the point of view of cooperative game theory. González and Herrero (2004) is one of the scarce papers in which cooperation is analyzed in queueing models. It considers a Markovian situation in which several agents maintaining their own servers agree to cooperate and hold a common server for their populations. Each agent has specified a maximum value for the expected time in the system of the members of his population. The problem of how to allocate among the agents the cost of a common server, that fulfills the specification of each one, is dealt with, and applied to a cost sharing problem in the Spanish health system.

The study of cooperation in queueing models is a relevant issue which deserves the attention of game theorists and operations researchers. In many real world situations several providers of a particular service agree to maintain common servers which are available for all their populations: think of a group of banks which share a network of cash machines, a cluster of universities which hold one high-performance computer, or a set of hospitals keeping a joint blood bank. In all these situations questions like how to allocate the cost of the common servers or when a group of service providers should cooperate are really relevant and should be approached from a scientific point of view. We devote this paper to deal with these questions in some Markovian models.

The organization of this paper is as follows. In Section 2 we set up our notation and analyze two variations of the model in González and Herrero (2004). In the first one, each agent has a specification for the maximum time in the system and for the probability that one of his customers spends more than this maximum. In the second variation, agents are interested in the time in the queue instead of in the time in the system. In Section 3 we consider a new variation which allows for preemptive priority schemes to decrease the total cost. In this context a rule for allocating the holding costs of the common server is introduced and axiomatically characterized. This rule can be easily computed and, moreover, provides core allocations.

2. Basic Markovian models

Consider a basic queueing system where customers arrive requiring a service, have to queue while the unique server is occupied, are selected from the queue by a certain discipline (i.e., a specification of the order in which they are selected), and leave the system after having been served. An $M/M/1$ model describes a system of this kind, when the arrivals occur according to a Poisson process with parameter λ (i.e., inter-arrival times are independent and identically distributed following an exponential distribution with mean $\frac{1}{\lambda}$), the service time follows an exponential distribution with mean $\frac{1}{\mu}$, and the queue discipline is FCFS (first to come, first to be served). The steady state condition for this system is $\lambda < \mu$. From now on we deal only with $M/M/1$ systems in steady state. We assume that the reader is familiar with elementary issues of Markovian queues, more precisely, with the model $M/M/1$. Anyway, we briefly recall whenever needed some features in connection with that model (which is treated in deep, for instance, in Gross and Harris (1998)).

Consider a situation in which n agents run n $M/M/1$ systems which provide a similar service. Each agent $i \in N = \{1,\ldots,n\}$ runs his own queue and provides the service to his own population, $\lambda_i > 0$, $\mu_i > \lambda_i$ denoting the parameters characterizing agent i's $M/M/1$ system. Besides, each agent i wants that the average time that his customers spend in the system does not exceed a certain maximum value $t_i \in (0, +\infty)$. Moreover, the cost of maintaining a server is supposed to be a linear function of its efficiency, measured by the inverse of its expected service time (which, according to the properties of the exponential distribution, turns out to be the expected number of potential service completions per time unit), i.e., $c(i) = k\mu_i$, for all $i \in N$. Generally game theory deals with solutions which are invariant to scale changes, so we assume without loss of generality that $k = 1$. Now, since agents want to minimize the cost and since the expected time of a customer i in such an $M/M/1$ system is known to be $(\mu_i - \lambda_i)^{-1}$, then

$$t_i = \frac{1}{\mu_i - \lambda_i}$$

and thus

$$c(i) = \mu_i = \frac{1}{t_i} + \lambda_i.$$

González and Herrero (2004) considered the following question. How is the new situation if some agents agree to maintain one common server to attend their customers? They assume that this unique server should assure that the average time of a customer in the system is the lowest of the maximum admissible values for all the agents that make the arrangement (notice that this includes a feasibility assumption which guarantees that it is possible to ensure the desired service rate at the common server). If we take S the coalition of these agents and denote $t^S := \min\{t_i : i \in S\}$ and $\lambda_S := \sum_{i\in S}\lambda_i$, the cost of the unique server is

$$c(S) = \frac{1}{t^S} + \lambda_S. \tag{1}$$

Notice that $\sum_{i\in S} c(i) \geqslant c(S)$ when $|S| > 1$, so sharing the server in this way leads to a cost reduction. Eq. (1) defines a cost TU-game (N, c). Remember that a cost TU-game is a pair (N, c), where N is a finite set of agents and c is the characteristic function, which assigns for every $S \subset N$ a real number $c(S)$ that indicates the cost of a particular project for the agents in coalition S. It is common to identify the game (N, c) with its characteristic function.

González and Herrero (2004) observe that c defined by (1) is the sum of an additive game plus an airport game. So, c is a concave game and its core is known to be the convex hull of the marginal contribution vectors. Moreover, its Shapley value $\Phi(c)$ can be easily computed and provides core allocations. For details on concave games and on airport games the reader can consult Owen (1995).

In the following we extend the model in (1) to deal not only with expected times. We consider the case where every agent i needs to guarantee for each of his customers that his time in the system will be smaller than or equal to a critical value ω_i with a sufficiently high-probability $1 - \alpha_i$. In this case the cost of the unique server for coalition S is given in the following proposition.

Proposition 1. *In the conditions above, the cost of a common server which fulfills the conditions of the agents in S is given by*:

$$\hat{c}(S) = \lambda_S + \max_{i\in S}\left\{\frac{-\ln \alpha_i}{\omega_i}\right\}. \tag{2}$$

Proof. Let us denote by $\mathscr{W}_i$ the random variable "time in the system spent by a customer of type i". Therefore, for all $i \in N$, the condition

$$P(\mathscr{W}_i \leqslant \omega_i) \geqslant 1 - \alpha_i,$$

must hold. It is a well-known result that the time that a customer spends in an $M/M/1$ system with parameters λ and μ follows an exponential distribution with mean $\frac{1}{\mu-\lambda}$. So, i will maintain a server with service rate μ_i such that

$$P(\mathscr{W}_i \leqslant \omega_i) = 1 - e^{-(\mu_i-\lambda_i)\omega_i} = 1 - \alpha_i,$$

which implies that

$$\ln \alpha_i = -(\mu_i - \lambda_i)\omega_i$$

and thus

$$\mu_i = \lambda_i - \frac{\ln \alpha_i}{\omega_i}.$$

Now if a coalition S forms to maintain a common server which fulfills the specifications of all the agents, it should be satisfied that, for all $i \in S$,

$$P(\mathscr{W}_S \leqslant \omega_i) \geqslant 1 - \alpha_i,$$

where $\mathscr{W}_S$ is the random variable "time in the system spent by a customer of any agent in S". Then, the service rate μ of the server must satisfy for every $i \in S$:

$$1 - e^{-\left(\mu-\sum_{i\in S}\lambda_i\right)\omega_i} \geqslant 1 - \alpha_i,$$

which implies that

$$\mu \geqslant \lambda_S - \frac{\ln \alpha_i}{\omega_i}$$

for all $i \in S$. So, for all $S \subset N$,

$$\hat{c}(S) = \lambda_S + \max_{i \in S} \left\{ \frac{-\ln \alpha_i}{\omega_i} \right\}. \qquad \square$$

Eq. (2) defines a cost TU-game $(N, \hat{c})$. We remark that our model includes as a particular case the cost game in González and Herrero (2004). Indeed, taking $\alpha_i = \frac{1}{e}$, for all $i \in N$, we obtain exactly the same game as in their paper. Moreover, we note again that $\hat{c}$ is the sum of an additive game plus an airport game which, once more, implies that $\hat{c}$ is concave, its core can be fully described and its Shapley value provides a specially noticeable core allocation. Following Littlechild and Owen (1973), the next corollary gives an explicit expression of the Shapley value in this context.

Corollary 1. *The Shapley value of the game $(N, \hat{c})$ is given by*

$$\Phi_{\pi(i)}(\hat{c}) = \frac{-\ln \alpha_{\pi(1)}}{n\omega_{\pi(1)}} + \frac{1}{n-1}\left(\frac{-\ln \alpha_{\pi(2)}}{\omega_{\pi(2)}} - \frac{-\ln \alpha_{\pi(1)}}{\omega_{\pi(1)}}\right) + \cdots + \frac{1}{n-i+1}\left(\frac{-\ln \alpha_{\pi(i)}}{\omega_{\pi(i)}} - \frac{-\ln \alpha_{\pi(i-1)}}{\omega_{\pi(i-1)}}\right) + \lambda_{\pi(i)},$$

for all $i \in N$, and where π is a permutation of N such that

$$\frac{-\ln \alpha_{\pi(1)}}{\omega_{\pi(1)}} \leqslant \frac{-\ln \alpha_{\pi(2)}}{\omega_{\pi(2)}} \leqslant \cdots \leqslant \frac{-\ln \alpha_{\pi(n)}}{\omega_{\pi(n)}}.$$

The rest of the section shows some difficulties when extending the previous results to the case in which the agents are concerned with expected times *in the queue* instead of in the system. Surprisingly enough, this slight variation leads to a scenario in which sometimes it is better for the agents not to cooperate.

Consider the same situation as in González and Herrero (2004) but in such a way that now each agent i has a maximum admissible value $t_i^q \in (0, +\infty)$ for the expected time of his customers in the queue. In a stationary $M/M/1$ system with parameters λ and μ, the average waiting time in the queue by a customer is given by

$$\frac{\lambda}{\mu(\mu - \lambda)}.$$

So, each i will choose a server with an expected service time μ_i such that

$$t_i^q \mu_i^2 - t_i^q \mu_i \lambda_i - \lambda_i = 0,$$

which implies that

$$\mu_i = \frac{t_i^q \lambda_i \pm \sqrt{(t_i^q)^2 \lambda_i^2 + 4\lambda_i t_i^q}}{2t_i^q} = \frac{\lambda_i}{2} \pm \sqrt{\left(\frac{\lambda_i}{2}\right)^2 + \frac{\lambda_i}{t_i^q}}.$$

From the fact that $\mu_i > \lambda_i$ for all $i \in N$, it follows that only the positive square root is possible, and rewriting the expression, the cost of maintaining a server i in this situation is

$$c^q(i) = \frac{\lambda_i}{2} + \frac{\lambda_i}{2}\sqrt{1 + \frac{4}{\lambda_i t_i^q}}.$$

If coalition $S \subset N$ forms,

$$c^q(S) = \frac{\lambda_S}{2} + \sqrt{\left(\frac{\lambda_S}{2}\right)^2 + \frac{\lambda_S}{t^{qS}}} = \frac{\lambda_S}{2} + \frac{\lambda_S}{2}\sqrt{1 + \frac{4}{\lambda_S t^{qS}}},$$

where $t^{qS} = \min_{i \in S}\{t_i^q\}$.

The following example shows that in a situation like this, players may prefer not to cooperate.

Example 1. Take $N = \{1,2\}$ and $\lambda_1 = \lambda_2 = 1$, $t_1^q = 100$, $t_2^q = 1$. Then:

- $c^q(N) = 1 + \sqrt{1 + \frac{2}{1}} = 1 + \sqrt{3}$.
- $c^q(1) + c^q(2) = \frac{1}{2} + \sqrt{(\frac{1}{2})^2 + \frac{1}{100}} + \frac{1}{2} + \sqrt{(\frac{1}{2})^2 + \frac{1}{1}} = 1 + \sqrt{0.26} + \sqrt{1.25}$.

Hence, $c^q(1) + c^q(2) < c^q(N)$.

So, in the case that the agents are concerned with the time their customers spend in the queue, instead of with the time their customers spend in the system, maybe they will not have incentives to cooperate, at least under the conditions considered up to now. Next proposition gives a sufficient condition that makes cooperation to be a good option.

Proposition 2. *A sufficient condition in order that* $\sum_{i\in S} c^q(i) \geqslant c^q(S)$ *for a coalition* $S \subset N$ *is that*

$$\lambda_i t_i^q \leqslant \lambda_S t^{qS} \tag{3}$$

for all $i \in S$.

The proof is immediate if the terms $\sum_{i\in S} \frac{\lambda_i}{2}\sqrt{1 + \frac{4}{\lambda_i t_i^q}}$ and $\frac{\lambda_S}{2}\sqrt{1 + \frac{4}{\lambda_S t^{qS}}}$ are compared. Taking $t_1^q = 4$ in the example above, one checks that the condition in Proposition 2 is not necessary.

The interpretation of condition (3) is clear. It says that the common server has to be able to take on more work than each one of the individual servers whilst maintaining expected sojourn time guarantees for the individual agents. In particular, this is true when the values t_i^q are homogeneous (notice that in Example 1 above t_1^q and t_2^q are strongly discrepant).

We finish this section with two remarks. The first has to do with the motivation of the two new models treated here. The second is a technical comment.

Remark 1. The two new models treated in this section are very natural and can be applied in many different scenarios, for instance in the cost sharing problem in the Spanish health system described in González and Herrero (2004). In fact, it is quite sensible to specify, for some specially delicate pathologies, a maximum value for the time in the system (with a high-probability) instead of a maximum value for the expected time in the system. In this context, it is also reasonable to deal with times in the queue instead of times in the system, because what we want to diminish are the waiting times for a surgical intervention instead of the time of the surgery itself.

Remark 2. In this section we have considered queueing systems with an FCFS discipline. Actually, this assumption is only necessary to obtain expression (2) for $\hat{c}$. The expressions for c and c^q given in this section are still valid if we simply assume that the system discipline satisfies the conservation law (see Kleinrock (1976) for details on the conservation law).

3. Cooperation under preemptive priority

In this section we deal with the following question. Taking into account that the different players have different specifications for their populations, would it be helpful in order to diminish the cost of the common server that a priority scheme in the queue discipline is adopted?

We assume that the agents in N have agreed to run a common server to attend their customers. However, now we suppose that a priority scheme with n classes (one for each agent) has been established. In this section we always deal with priority schemes allowing preemption. Each class $i \in N$ corresponds to agent i, so it generates an expected number of clients per time unit λ_i, and it has a maximum value t_i for the expected waiting time in the system. We will moreover allow the use of *mixing* priority schemes. A mixing priority scheme (or *priority policy*) consists of multiplexing a finite set of priority schemes in such a way that each of them will operate during a desired percentage of time. The following theorem proves that in this context there always exists a priority policy whose associated cost is less than or equal to the cost of the FCFS system given in (1).

Theorem 1. *For any vector $(t_1,\dots,t_n)$ of maximum expected waiting times in the system for the agents in N, there exists a priority policy that ensures these waiting times with a cost less than or equal to the one given by the approach in (1).*

Proof. Let us denote by $\Pi(N)$ the set of permutations of the finite set N. Let $\sigma\in\Pi(N)$ be an ordering of the n classes which establishes the priority scheme of the queue. Here $\sigma(i)$ represents the position which has been assigned to the class i. The smaller the position index, the higher priority associated to the class. It is well-known (see, for instance Gross and Harris (1998, p. 233)) that for any $\mu>\lambda_N$, the expected waiting time in the system for each class i under the priority scheme σ is

$$W_i(\sigma,\mu)=\frac{\mu}{\left(\mu-\sum_{j:\sigma(j)<\sigma(i)}\lambda_j\right)\left(\mu-\sum_{j:\sigma(j)\leqslant\sigma(i)}\lambda_j\right)}. \tag{4}$$

Notice that $W_i(\sigma,\mu)$ is a decreasing function of μ. We denote by $W(\sigma,\mu)$ the vector whose coordinates are given by (4) and by $\mathscr{F}(N,\mu)$ the set

$$\mathscr{F}(N,\mu)=\operatorname{conv}\{W(\sigma,\mu)\in\mathbb{R}^n:\ \sigma\in\Pi(N)\},$$

where conv stands for convex hull. We distinguish two cases.

Case 1. $t\in\mathscr{F}(N,\mu)$ for some $\mu>\lambda_N$.

Theorem 2 in Coffman and Mitrani (1980) established that $t=(t_1,\dots,t_n)$ is achievable by some priority policy using a common server with a service rate μ if and only if $(t_1,\dots,t_n)\in\mathscr{F}(N,\mu)$. Then, since t, λ and μ must satisfy the conservation law for queueing disciplines (see, e.g., Kleinrock (1976, p. 114)) the following equation holds:

$$\sum_{i\in N}\frac{\lambda_i}{\lambda_N}t_i=\frac{1}{\mu-\lambda_N}. \tag{5}$$

Hence, the common service rate μ can be obtained solving Eq. (5). Its value is

$$\mu=\lambda_N+\frac{\lambda_N}{\sum_{i=1}^n\lambda_i t_i}. \tag{6}$$

Clearly,

$$\mu\leqslant\lambda_N+\frac{1}{t^N}, \tag{7}$$

so, in view of (1), the cost of the common server diminishes if a priority scheme is adopted under which the required vector $(t_1,\dots,t_n)$ is in $\mathscr{F}(N,\mu)$.

Case 2. $t\notin\bigcup_{\mu>\lambda_N}\mathscr{F}(N,\mu)$.

From Lemma 2 in Coffman and Mitrani (1980), it is derived that any vector of expected waiting times in the system $(t_1,\dots,t_n)$ with $t_i=t_j$ for every $i,j\in N$ belongs to the interior of $\mathscr{F}(N,\mu)$ for some μ. (Note that, in this case, $\mu=\lambda_N+\frac{1}{t^N}$.)

To prove the result, assume without loss of generality that $t_1=t^N$. Let $l(t)$ be the line segment with extreme points t and $\hat{t}=(t_1,\dots,t_1)$. The segment $l(t)$ is included in the halfspace $H^+=\left\{x\in\mathbb{R}^n:\sum_{i\in N}\frac{\lambda_i}{\lambda_N}x_i\geqslant t^N\right\}$. Indeed, the hyperplane defining the halfspace H^+ contains $\hat{t}$ and its normal vector $\left(\frac{\lambda_1}{\lambda_N},\dots,\frac{\lambda_n}{\lambda_N}\right)\geqslant 0$. Thus, $\hat{t}+\mathbb{R}^n_+\subset H^+$. Now, since clearly $l(t)\subset\hat{t}+\mathbb{R}^n_+$, the inclusion $l(t)\subset H^+$ follows.

The above construction proves that $l(t)$ intersects $\bigcup_{\mu>\lambda_N}\mathscr{F}(N,\mu)$ in a subsegment. All the points in that intersection, with the exception of $\hat{t}$, are attainable by priority policies with service rates smaller than $\frac{1}{t^N}+\lambda_N$, that corresponds to the policy attaining $\hat{t}$. (Notice that the service rate decreases while $\|W\|$ increases along the ray $\{x\in\mathbb{R}^n_+:x_1=x_2=\cdots=x_n>0\}$, see (4). An illustration can be found in Fig. 2.)

Hence, any service rate μ^* associated with a point

$$t^* \in (l(t) \setminus \{\hat{t}\}) \cap \bigcup_{\mu > \lambda_N} \mathscr{F}(N, \mu),$$

satisfies the aspiration level given by t and with a service rate smaller than the one in (1), namely $\frac{1}{t^N} + \lambda_N$. □

Now we illustrate the result above for the two-classes situation. Here, the extreme points of the set $\mathscr{F}(N, \mu)$ are given by

$$\left(\frac{1}{\mu - \lambda_1}, \frac{\mu}{(\mu - \lambda_1)(\mu - \lambda_N)}\right), \quad \left(\frac{\mu}{(\mu - \lambda_2)(\mu - \lambda_N)}, \frac{1}{\mu - \lambda_2}\right).$$

Of course, it must hold that $\mu > \lambda_N = \lambda_1 + \lambda_2$. Fig. 1 illustrates this result where $\lambda_1 = \lambda_2 = 1$.

According to Theorem 2 in Coffman and Mitrani (1980), any (t_1, t_2) which lies inside the region limited by the curves corresponding to the orderings σ and τ is achievable using a certain priority policy, by a server with common service rate

$$\mu = \lambda_1 + \lambda_2 + \frac{\lambda_1 + \lambda_2}{\lambda_1 t_1 + \lambda_2 t_2},$$

as it is derived from the conservation law (5).

Fig. 2 displays the case in which $(t_1, t_2) \notin \bigcup_{\mu > \lambda_N} \mathscr{F}(N, \mu)$. The cost associated with $\bar{t}$, according to Theorem 1, is less than or equal to $\frac{1}{t^N} + \lambda_N$.

From now on we consider problems where the expected waiting time vector in the system $t = (t_i)_{i \in N}$ is achievable. (Notice that t being achievable implies that for any $S \subset N$ then $(t_i)_{i \in S}$ is achievable as well.) Let us denote by QS the set of queueing situations $(N, \{\lambda_i\}_{i \in N}, \{t_i\}_{i \in N})$ such that N is finite and $(t_i)_{i \in N} \in \mathscr{F}(N, \mu)$ with $\mu = \lambda_N + \frac{\lambda_N}{\sum_{i \in N} \lambda_i t_i}$. Then, the maintenance cost of a common server for any coalition $S \subset N$ is given by

$$\bar{c}(S) = \lambda_S + \frac{\lambda_S}{\sum_{i \in S} \lambda_i t_i}. \tag{8}$$

Notice that, as we have already remarked, $\bar{c}(N)$ is smaller than or equal to the total cost in González and Herrero's model.

The problem now is how to allocate $\bar{c}(N)$ among the agents. In order to do it, we consider the cost TU-game $(N, \bar{c})$ given by Eq. (8). Observe first that, for each $S \subset N$,

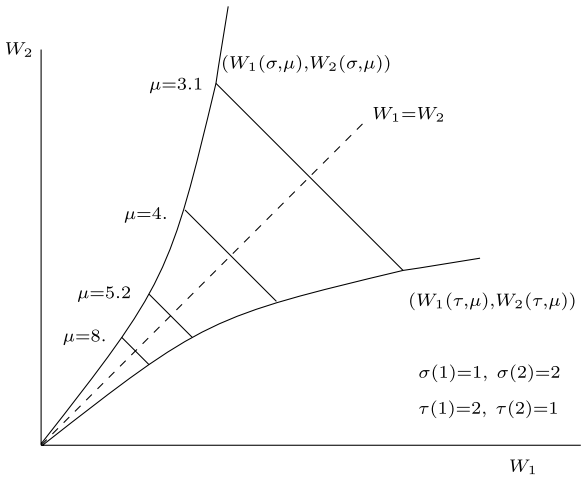

Fig. 1. Vectors of achievable expected waiting times in the system.

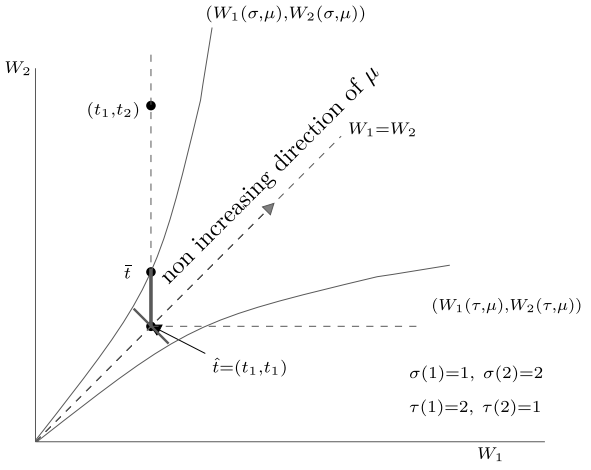

Fig. 2. The case in which $(t_1,t_2)\notin\cup_{\mu>\lambda_N}\mathscr{F}(N,\mu)$.

$$\bar{c}(S)-\sum_{i\in S}\bar{c}(\{i\})=\bar{c}(S)-\sum_{i\in S}\left(\lambda_i+\frac{1}{t_i}\right)\leqslant 0.$$

In the class of queueing situations QS, an allocation rule f is a function which associates to each $(N,\{\lambda_i\}_{i\in N},\{t_i\}_{i\in N})\in$ QS a non-negative vector in $\mathbb{R}^N$, denoted by $f(N,\{\lambda_i\}_{i\in N},\{t_i\}_{i\in N})$, such that the sum of its components equals $\bar{c}(N)$. We define the *proportional allocation rule*, denoted by φ^p, as

$$\varphi_i^p\left(N,\{\lambda_i\}_{i\in N},\{t_i\}_{i\in N}\right)=\lambda_i+\frac{\lambda_i}{\sum_{j\in N}\lambda_j t_j}.$$

According to this rule, each agent $i\in N$ pays an additive part λ_i plus a splitting of $\frac{\lambda_N}{\sum_{j\in N}\lambda_j t_j}$ proportional to λ_i. Notice that

$$\lambda_i+\frac{\lambda_i}{\sum_{j\in N}\lambda_j t_j}=\frac{\lambda_i}{\lambda_N}\bar{c}(N).$$

So φ^p can also be said to allocate to each agent a splitting of $\bar{c}(N)$ proportional to λ_i.

An important property for an allocation rule f is that it provides *core allocations*. In this context this means that, for every $(N,\{\lambda_i\}_{i\in N},\{t_i\}_{i\in N})\in$ QS and every $S\subset N$,

$$\sum_{i\in S}f_i\left(N,\{\lambda_i\}_{i\in N},\{t_i\}_{i\in N}\right)\leqslant\bar{c}(S),$$

or, in words, that the allocation of the total cost $\bar{c}(N)$ is acceptable for every coalition $S\subset N$. The following proposition shows that the proportional allocation rule in fact provides core allocations.

Proposition 3. *φ^p provides core allocations.*

Proof. For each coalition $S\neq\emptyset$, the difference $\sum_{i\in S}\varphi_i^p\left(N,\{\lambda_i\}_{i\in N},\{t_i\}_{i\in N}\right)-\bar{c}(S)$ is

$$\sum_{i\in S}\frac{\lambda_i}{\sum_{j\in N}\lambda_j t_j}-\sum_{i\in S}\frac{\lambda_i}{\sum_{j\in S}\lambda_j t_j}\leqslant 0. \qquad \square$$

An obvious consequence of Proposition 3 is that each game $\bar{c}$ associated with an element of QS is totally balanced. The following example shows that $\bar{c}$ needs not to be concave and also that the Shapley value of $\bar{c}$ may fall outside its core.

Example 2. Take $N=\{1,2,3\}$, $\lambda_1=\lambda_2=\lambda_3=1$, and $t_1=1$, $t_2=47.29$ and $t_3=53.71$. After some algebra, it is easy to check that $(N,\{\lambda_i\}_{i\in N},\{t_i\}_{i\in N})\in$ QS. Then we have:

- $\bar{c}(\{1\}) = 2$, $\bar{c}(\{2\}) = 1.021$, $\bar{c}(\{3\}) = 1.019$,
- $\bar{c}(\{1,2\}) = 2.041$, $\bar{c}(\{1,3\}) = 2.037$, $\bar{c}(\{2,3\}) = 2.02$,
- $\bar{c}(\{1,2,3\}) = 3.029$.

Consider $S = \{1\} \subset T = \{1,2\}$ and $i = 3$. Then

$$\bar{c}(S \cup \{i\}) - \bar{c}(S) < \bar{c}(T \cup \{i\}) - \bar{c}(T),$$

so $\bar{c}$ is not a concave game. Moreover, the Shapley value of this game is $\Phi(\bar{c}) = (1.343, 0.845, 0.841)$, which is not a core allocation because $\Phi_1(\bar{c}) + \Phi_2(\bar{c}) = 2.188 > \bar{c}(\{1,2\}) = 2.041$.

In summary, φ^p is a reasonable allocation rule that (a) can be easily computed and (b) provides core allocations. So, this rule is our proposal for allocating the maintenance cost of the common server in this context. We finish the paper providing an axiomatic characterization of this rule which shows that it has excellent properties from the point of view of the immunity to possible manipulations.

To start with, let us introduce two appealing properties for an allocation rule f defined on QS.

P1. Non-advantageous reallocation

Let $(N, \{\lambda_i\}_{i\in N}, \{t_i\}_{i\in N}) \in$ QS and $(N, \{\tilde{\lambda}_i\}_{i\in N}, \{\tilde{t}_i\}_{i\in N}) \in$ QS be such that $\sum_{i\in N}\lambda_i t_i = \sum_{i\in N}\tilde{\lambda}_i\tilde{t}_i$ and $\lambda_N = \tilde{\lambda}_N$. Then

$$\sum_{i\in T} f_i\big(N, \{\lambda_i\}_{i\in N}, \{t_i\}_{i\in N}\big) = \sum_{i\in T} f_i\Big(N, \{\tilde{\lambda}_i\}_{i\in N}, \{\tilde{t}_i\}_{i\in N}\Big)$$

for any $T \subset N$ with $\lambda_T = \tilde{\lambda}_T$.

The meaning of this property is that a rule should be invariant to reallocations of the parameters λ_i within any coalition T while keeping the total cost. This reallocation is one possible way in which a certain coalition can manipulate its parameters to obtain some advantage. Another possible way is performing artificial mergings or splittings. These manipulations are prevented by the next property. Before its introduction we need the following definition.

Definition 1. Let $(N, \{\lambda_i\}_{i\in N}, \{t_i\}_{i\in N}) \in$ QS be such that $t_i = t$, for every $i \in N$. Then for each $S \subset N$, the S-manipulation of $(N, \{\lambda_i\}_{i\in N}, \{t_i\}_{i\in N})$ is the queueing situation $(N^S, \{\lambda_i\}_{i\in N^S}, \{t_i\}_{i\in N^S}) \in$ QS where

- $N^S = (N \backslash S) \cup \{i_S\}$,
- $\lambda_{i_S} = \sum_{i\in S}\lambda_i$, and
- $t_{i_S} = t$.

Notice that, in these conditions, $\bar{c}(N) = \bar{c}(N^S)$ for every $S \subset N$. Now we present the second property.

P2. Non-advantageous merging or splitting

Let $(N, \{\lambda_i\}_{i\in N}, \{t_i\}_{i\in N}) \in$ QS be such that $t_i = t$, for every $i \in N$. Then, for each $S \subset N$,

$$f_{i_S}(N^S, \{\lambda_i\}_{i\in N^S}, \{t_i\}_{i\in N^S}) = \sum_{i\in S} f_i(N, \{\lambda_i\}_{i\in N}, \{t_i\}_{i\in N}).$$

It is clear that the proportional allocation rule φ^p satisfies P1 and P2. Moreover, the next theorem shows that these two properties characterize the proportional allocation rule.

Theorem 2. *The proportional allocation rule φ^p is the unique allocation rule defined on* QS *which satisfies P1 and P2.*

Proof. We have already mentioned that φ^p satisfies P1 and P2. Let us check its uniqueness. Take an allocation rule f defined on QS which satisfies P1 and P2. Consider $(N, \{\lambda_i\}_{i\in N}, \{t_i\}_{i\in\ N}) \in$ QS and fix arbitrarily $j \in N$. We have to prove that

$$f_j\big(N,\{\lambda_i\}_{i\in N},\{t_i\}_{i\in N}\big)=\lambda_j+\frac{\lambda_j}{\sum_{i\in N}\lambda_i t_i}.$$

We define for all $i\in N$

$$\tilde{t}_i=\sum_{k\in N}\frac{\lambda_k}{\lambda_N}t_k=\tilde{t};$$

observe that P1 implies that

$$f_j\big(N,\{\lambda_i\}_{i\in N},\{t_i\}_{i\in N}\big)=f_j\big(N,\{\lambda_i\}_{i\in N},\{\tilde{t}_i\}_{i\in N}\big), \tag{9}$$

(notice that P1 can be applied because $\sum_{i\in N}\lambda_i\tilde{t}_i=\sum_{i\in N}\lambda_i t_i$).

Now take the S-manipulation of $(N,\{\lambda_i\}_{i\in N},\{\tilde{t}_i\}_{i\in N})$ for $S=N\setminus\{j\}$. We know that:

- $\bar{c}(N^S)=\bar{c}(N)$, (10)
- $\bar{c}(N^S)=f_j(N^S,\{\lambda_i\}_{i\in N^S},\{\tilde{t}_i\}_{i\in N^S})+f_{is}(N^S,\{\lambda_i\}_{i\in N^S},\{\tilde{t}_i\}_{i\in N^S})$, (11)
- $\bar{c}(N)=f_j(N,\{\lambda_i\}_{i\in N},\{\tilde{t}_i\}_{i\in N})+\sum_{k\neq j}f_k(N,\{\lambda_i\}_{i\in N},\{\tilde{t}_i\}_{i\in N})$. (12)

Since (10)–(12) hold, and f satisfies P2, then

$$f_j(N,\{\lambda_i\}_{i\in N},\{\tilde{t}_i\}_{i\in N})=f_j(N^S,\{\lambda_i\}_{i\in N^S},\{\tilde{t}_i\}_{i\in N^S}).$$

Hence, it is clear that $f_j(N,\{\lambda_i\}_{i\in N},\{\tilde{t}_i\}_{i\in N})$ can be written as a function of λ_N, λ_j and $\tilde{t}_j$ (notice that $\tilde{t}_j=\tilde{t}$ does not really depend on j). So,

$$f_j(N,\{\lambda_i\}_{i\in N},\{\tilde{t}_i\}_{i\in N})=F(\lambda_N,\lambda_j,\tilde{t}).$$

Suppose that F is linear in its second variable (we will prove below that this is actually true). Then,

$$f_j(N,\{\lambda_i\}_{i\in N},\{\tilde{t}_i\}_{i\in N})=g(\lambda_N,\tilde{t})\lambda_j. \tag{13}$$

Thus

$$\bar{c}(N)=\sum_{j\in N}f_j(N,\{\lambda_i\}_{i\in N},\{\tilde{t}_i\}_{i\in N})=g(\lambda_N,\tilde{t})\lambda_N,$$

and so $g(\lambda_N,\tilde{t})=\frac{\bar{c}(N)}{\lambda_N}$. Now, in view of (9) and (13), we get:

$$f_j(N,\{\lambda_i\}_{i\in N},\{t_i\}_{i\in N})=\frac{\bar{c}(N)}{\lambda_N}\lambda_j=\lambda_j+\frac{\lambda_j}{\sum_{i\in N}\lambda_i t_i}.$$

So, to finish the proof we just need to check that F is linear in its second variable. Notice that we have a collection of functions

$$\{F(\alpha,\cdot,\beta)|\alpha,\beta\in(0,+\infty)\}.$$

such that $F(\alpha,\cdot,\beta):(0,\alpha]\to\left[0,\alpha+\frac{1}{\beta}\right]$, for all $\alpha,\beta\in(0,+\infty)$. Let us take now α, β, x, $y\in(0,+\infty)$ with $x+y\leqslant\alpha$. Then, there exists $(N,\{\lambda_i\}_{i\in N},\{t_i\}_{i\in N})\in$ QS where $t_i=\beta$ for every $i\in N$, $\lambda_1=x$, and $\lambda_2=y$. Define the S-manipulation of this problem for $S=\{1,2\}$. Then, since f satisfies P2,

$$\begin{aligned}F(\alpha,x+y,\beta)&=f_{is}\big(N^S,\{\lambda_i\}_{i\in N^S},\{t_i\}_{i\in N^S}\big)=f_1\big(N,\{\lambda_i\}_{i\in N},\{t_i\}_{i\in N}\big)+f_2\big(N,\{\lambda_i\}_{i\in N},\{t_i\}_{i\in N}\big)\\&=F(\alpha,x,\beta)+F(\alpha,y,\beta).\end{aligned}$$

So, for every α, $\beta\in(0,+\infty)$, $F(\alpha,\cdot,\beta)$ is additive. Since $F(\alpha,\cdot,\beta)$ is also non-negative, it is clear that it is increasing. It is an easy exercise to prove that every additive, increasing function $h:(0,\alpha]\to\left[0,\alpha+\frac{1}{\beta}\right]$ is also linear. This completes the proof. □

Finally we check that these two properties are independent.

1. The rule f^1 which assigns to each queueing situation $(N, \{\lambda_i\}_{i\in N}, \{t_i\}_{i\in N}) \in QS$ the vector whose jth coordinate is given by

$$f_j^1\left(N, \{\lambda_i\}_{i\in N}, \{t_i\}_{i\in N}\right) = \frac{\bar{c}(N)}{|N|},$$

where $|N|$ is the number of agents in N, satisfies P1, but it does not satisfy P2.

2. The rule f^2 which assigns to each queueing situation $(N, \{\lambda_i\}_{i\in N}, \{t_i\}_{i\in N}) \in QS$ the vector whose jth coordinate is given by

$$f_j^2(N, \{\lambda_i\}_{i\in N}, \{t_i\}_{i\in N}) = \frac{\lambda_j t_j}{\sum_{i\in N}\lambda_i t_i}\bar{c}(N)$$

satisfies P2, but it does not satisfy P1.

References

Borm, P., Hamers, H., Hendrickx, R., 2001. Operations research games: A survey. Top 9, 139–216.

Coffman, E.G., Mitrani, I., 1980. A characterization of waiting time performance realizable by single-server queues. Operations Research 28, 810–821.

Curiel, I., Hamers, H., Klijn, F., 2002. Sequencing games: A survey. In: Borm, P., Peters, H. (Eds.), Chapters in Game Theory. Kluwer Academic Publishers, pp. 27–50.

González, P., Herrero, C., 2004. Optimal sharing of surgical costs in the presence of queues. Mathematical Methods of Operations Research 59, 435–446.

Gross, D., Harris, C.M., 1998. Fundamentals of queueing theory. Wiley.

Hassin, R., Haviv, M., 2003. To queue or not to queue. Kluwer Academic Publishers.

Kleinrock, L., 1976. Queueing Systems, Volume II: Computer Applications. Wiley.

Littlechild, S.C., Owen, G., 1973. A simple expression for the Shapley value in a special case. Management Science 20, 370–372.

Maniquet, F., 2003. A characterization of the Shapley value in queueing problems. Journal of Economic Theory 109, 90–103.

Moulin, H., Stong, R., 2002. Fair queuing and other probabilistic allocation methods. Mathematics of Operations Research 27, 1–30.

Owen, G., 1995. Game Theory. Academic Press.

Tadj, L., Choudhury, G., 2005. Optimal design and control of queues. Top 13, 359–412.