European Journal of Operational Research 000 (2018) 1–8

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Stochastics and Statistics

Cooperation and sharing costs in a tandem queueing network

Dan Bendel, Moshe Haviv*

Department of Statistics and the Fedremann Center for the Study of Rationality, Hebrew University of Jerusalem, Jerusalem 91905, Israel

ARTICLE INFO

Article history:
Received 11 January 2017
Accepted 29 April 2018
Available online xxx

Keywords:
Queueing
Tandem queues
Cooperative games
Cost allocation
Core
Shapley value

ABSTRACT

We consider a tandem network of queues with a Poisson arrival process to the first queue. Service times are assumed to be exponential. In cases where they are not, we additionally assume a processor sharing service discipline in all servers. Consecutive servers may cooperate by pooling resources which leads to the formation of a single combined server that satisfies the aggregated service demands with a greater service rate. On this basis we define a cooperative game with transferable utility, where the cost of a coalition is the steady-state mean total number of customers in the system formed by its members. We show that the game is subadditive, leading to full cooperation being socially optimal. We then show the non-emptiness of the core, despite the characteristic function being neither monotone, nor concave. Finally, we derive several well-known solution concepts, including the Shapley value, the Banzhaf, value and the nucleolus, for the case where servers have equal mean service demands. In particular, we show that all three values coincide in this case.

1. Introduction

Cooperation among servers may lead to a reduction in overall congestion, compared with the case where they function individually. The reduction level is of course determined by the technology involved, which may limit the level of cooperation. However, from the servers' point of view, it is exogenous and hence beyond the capacity or control of the decision makers, i.e., the servers or their owners. What is not out of their hands, though, is the decision on how to split the gains due to cooperation among the participants so as to guarantee stability, namely, to make all of them somehow happy and eliminate their desire to act individually or to form smaller cliques who cooperate only among themselves. The question of which splits lead to stability is an important one and this is where cooperative game theory plays an important role.

The essence of cooperative game theory is to define a function, called the *characteristic function*, that for each subset of players (in our case servers), called a *coalition*, returns its cost. Then various solutions concepts, such as the Shapley value or the Banzhaf value, get this characteristic function as their input, and compute a numerical value for each player, stating her "fair" cost of participation. In our queueing application, the characteristic function states for each coalition what is the mean number of customers in the combined system (in particular, in the grand coalition, when all servers join forces), while the solution then assigns some cost to each of the servers.

In Section 2 below we review some of the existing literature. What all the models dealt with in the literature so far have in common is that different types of cooperation lead to different reductions in the congestion level. For example, in the model dealt with in Anily and Haviv (2010), cooperating servers form a single server that has a single arrival stream, where the corresponding arrival and service rates are the sums of the individual servers. Another model is dealt with in Timmer and Scheinhardt (2010, 2013). There, servers that cooperate still serve their original customers but they optimize the total congestion level by redistributing the total service capacity among individual servers. Finally, in the counterpart model dealt with in Anily and Haviv (2017), servers that cooperate maintain their service rates and optimization is now with respect to splitting the total arrival stream among the servers.

We deal here with another type of cooperation among servers, which, apparently for the first time, involves servers that are located in some physical topology, in our case a line, and cooperation is limited only among servers that are geographically close or, more precisely, that appear consecutively along the line. When, say, s consecutive identical servers cooperate they are functioning as a single server that serves the sum of the individual service demand but at a rate that is the sum of the individual rates. Such cooperation leads to a reduction in the sum of the queue lengths and it is natural to agree that full cooperation will be achieved, leading to a reduced, indeed, optimal, queue length. The next question is how the total costs should be divided among the servers.

* Corresponding author.
E-mail address: haviv@mscc.huji.ac.il (M. Haviv).

https://doi.org/10.1016/j.ejor.2018.04.049
0377-2217/

Our main result shows that this game is totally balanced; that is, this game and its subgames possess a core allocation. By core allocation we mean allocating the cost of the grand coalition in a way that no subset of players can cooperate among themselves in such a way that their new total cost will be less than the sum of the individual costs assigned to them. We would like to note that this result is established in spite of the fact (to be proved) that the game is neither monotone nor concave. We then deal with the special case where all servers are identical. We show that in this case the Shapley value, the Banzhaf value, and the nucleolus coincide. In particular, all three solution concepts assign half the cost to each of the two servers in the edges and nothing to each of the others. Finally, we show that in this special case the game is monotone non-decreasing and concave. What is novel in our results is that the cost assigned to the servers is not only a function of the individual arrival and server rates, but also a function of the index of the servers, reflecting their relative positions in the series of tandem queues.

2. Literature review

Many operations research problems, e.g., scheduling and production, have been studied in the context of cooperative game theory. Borm, Hamers, Hendrickx (2001) provides an extensive survey. Although there is a body of literature on pooling resources in queueing networks that assumes a central planner, e.g. Andradóttir and Ayhan (2005) and Argon and Andradóttir (2006), the research on the cooperation of independent servers is not abundant. Existing articles in the field consider variations of queueing models with independent service providers and deal with questions of cooperation profitability and cost sharing.

Anily and Haviv (2010) considers a model that consists of several M/M/1 queues with individual streams of customers. Servers may cooperate by merging their customers and pool capacities into a single M/M/1 queue. The cost of a coalition is the mean number of customers in the combined server formed by its members. It is shown that the game is totally balanced, and characterization of the non-negative part of the core is given. It is also shown that, unless all individual queues are identical, there is always a core allocation with at least one negative entry. Timmer and Scheinhardt (2010, 2013), consider single-server queues that preserve their autonomy but may redistribute their service capacities so as to optimize the total mean number of customers in the system. Core allocations are provided. A similar model, except that the individual service rate is preserved and customers are redirected so as to optimize a common objective, is stated in Anily and Haviv (2017). Here too a core allocation is identified.

Karsten, Slikker, and van Houtum (2012, 2014), and Özen, Reiman, and Wang (2011) consider the multi-server loss system where waiting is not possible and an arrival who finds all servers busy is lost. They study pooling of Erlang loss systems, where customers are redirected if the system is full, and prove the existence of a core allocation. In Karsten et al. (2012, 2014) results are derived by using extensions to the Erlang loss functions, whereas in Özen et al. (2011) the proofs are based on elasticity properties of the cost function. A similar model is later studied in Karsten, Slikker, and van Houtum (2015), where waiting in queues is possible. For an example of a production and congestion function model see (Yu, Benjaafar, & Gerchak, 2015).

See also Curiel, Pederzoli, and Tijs (1989) for analysis on sequencing games, where customers with cost functions depending on their completion time are facing a single server. Customers may find an optimal service order so as to minimize the total waiting costs. Gerichhausen and Hamers (2009) continues this line of research, introducing partitioning sequencing games, where agents arrive in batches. It is shown in the two above mentioned papers, that these games are convex, and an allocation rule that belongs to the core is specified. Moreover, a game independent expression for the Shapley value is presented.

3. Preliminaries on transferable utility cooperative games

A *cooperative game with transferable utility* is a pair (N, c) where

1. $N = \{1, \ldots, n\}$ is the set of players
2. $c : 2^N \to \mathbb{R}$ is a characteristic function with $c(\emptyset) = 0$

A subset S, $S \subseteq N$, is called a *coalition* and N is referred to as the *grand coalition*. The interpretation of $c(S)$, $S \subseteq N$, is the total cost incurred by the set of players S when cooperating and acting together as a group. A game (N, c) is called *monotone* non-decreasing if for $S \subseteq T$, $c(S) \le c(T)$. A game (N, c) is called *subadditive* if for S, $T \subset N$ with $S \cap T = \emptyset$, $c(S \cup T) \le c(S) + c(T)$. Note that cooperation in general, and the formation of the grand coalition in particular, is called for in subadditive games. Finally, a game (N, c) is said to be *concave* if for any two coalitions S, T such that $S \subset T \subset N$ and any player $i \in N \setminus T$,

$$c(S \cup \{i\}) - c(S) \ge c(T \cup \{i\}) - c(T). \tag{1}$$

We note that $c(S \cup \{i\}) - c(S)$, for $i \notin S$, is called the *marginal contribution* of player i to coalition S. Equivalently, the game is concave if for two coalitions S, $T \subseteq N$,

$$c(S \cup T) + c(S \cap T) \le c(S) + c(T). \tag{2}$$

Note that a concave game is subadditive but the converse is not necessarily true.

4. The model

Consider n servers who are placed one after another in a line. There exists a Poisson arrival process to server number 1. After service completion at server i, the customer moves to receive service from server $i+1$, $1 \le i \le n-1$. The customer leaves the system for good after receiving service from server n. The mean service demand at server i is denoted by $\bar{x}_i$, where $\bar{x}_i > 0$, $1 \le i \le n$[1]. For stability, we assume that $\lambda < \min_{i=1}^{n} \frac{1}{\bar{x}_i}$. We assume that (at least) one of the following conditions holds: (1) service times follow exponential distribution. In this case we do not assume anything further regarding the queue regime at each of the servers (as long as it is non-anticipating and work-conserving).[2] For example, it can be first-come first-served (FCFS). (2) There is no restriction on the service distribution, but the service regime belongs to the set of those that possess the $M \Rightarrow M$ property, namely, those that imply Poisson output in the case of Poisson input. The best known among them is the *processor sharing* (PS) regime. Under this regime, the server splits her capacity evenly among all those who seek service from her at any given moment (sometimes referred to as the *egalitarian processor sharing* (EPS) regime to reflect the assumption that no discrimination exists among all those who are served). In the case of standard exponential service (when all are getting the same product such as a cooked hamburger), PS is equivalent to the regime in which the one who is to receive the just-finished product is determined by a random lottery among all those present upon service completion. For more regimes that possess the $M \Rightarrow M$ property, see, e.g., Chapter 11 in Haviv (2013).

[1] There is a hidden assumption that all servers work at the same rate. But this assumption is without loss of generality (where service requirements are scaled accordingly).

[2] By "non-anticipating" we mean that the decision as to who receives service is not determined by the actual service requirements of those present in line. By "work-conserving" we mean that the total amount of work left in the system is the same as under a first-come first-served regime.

Please cite this article as: D. Bendel, M. Haviv, Cooperation and sharing costs in a tandem queueing network, European Journal of Operational Research (2018), https://doi.org/10.1016/j.ejor.2018.04.049