Memorandum 2060 (March 2017). ISSN 1874-4850. Available from: http://www.math.utwente.nl/publications
Department of Applied Mathematics, University of Twente, Enschede, The Netherlands

Customer and cost sharing in a Jackson network

Judith Timmer* Werner Scheinhardt*

February 7, 2017

Abstract

In this paper we consider a Jackson network of independent service stations. To improve upon the total expected waiting time, the stations may redistribute the total arrival rate of customers over all queues. We investigate this situation by means of arrival rate games. These are cooperative TU cost games in which the stations are the players and the total cost of a coalition is linear in the total expected waiting time. We show that these games are totally balanced; cooperation is beneficial to all stations. Further, we provide a tailor-made cost allocation rule that belongs to the core of the game.

1 Introduction

Consider a system of heterogeneous computing and communication resources. Besides the differences in service speeds, the job arrivals for the resources may be unequal. Therefore, the workload on the resources, or computers, may vary greatly. Improving the performance of the system by means of a better distribution of the work load is called load balancing [3].

In this paper we consider the work load in a Jackson network of independent service stations. Each station has a single queue with an individual service rate, and its own arrival rate of jobs or customers. The stations may redistribute their work loads, meaning that they reallocate their total arrival rate statically (i.e., not state-dependent). Our goal is to improve the total expected waiting time over the stations in the network.

Our problem is related to literature on load balancing for distributed systems. There, often cooperative games with non-transferable utility (NTU) are used to model the cooperation among the computers [1, 3, 8].

Also cooperative games with transferable utility (TU) are used. In [5] the core of cooperative queueing games is studied. The authors show that any single-attribute game embedded in an elastic function has a non-empty core. In [4] independent service providers are considered that may cooperate by pooling their resources and customer streams. The authors provide sufficient conditions for the corresponding cooperative TU game to have a core allocation.

*Stochastic Operations Research, Department of Applied Mathematics, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands. E-mail: {j.b.timmer, w.r.w.scheinhardt}@utwente.nl

In previous papers [9, 10] we analysed a Jackson network with independent service stations, in which the stations may redistribute their service rates to improve the total expected waiting time. It is shown that a tailor-made cost allocation always belongs to the core of the corresponding cooperative TU game.

In this paper we model the load balancing problem, concerning the *arrival rates* of the stations instead of their service rates, by means of cooperative TU games. We optimize the total expected waiting time of stations in the system, and provide an algorithm to calculate this optimum. Furthermore, we introduce a corresponding TU game, the *arrival rate game*, in which the stations are the players. Such a game is shown to be totally balanced, meaning that cooperation is beneficial to all stations. For this game we introduce a tailor-made cost allocation. This allocation is shown to belong to the core of the game.

2 Model and preliminaries

Consider a Jackson network with a set N of service stations. Station $i \in N$ has Poisson arrival rate λ_i and exponential service rate μ_i, $\mu_i > \lambda_i$. We refer to such a Jackson network as $(N, \{\mu_i\}, \{\lambda_i\})$. Without loss of generality assume $\mu_1 \geq \mu_2 \ldots \geq \mu_n$ with $n = |N|$. Due to these assumptions, the expected waiting time in station i (without cooperation) is given by $\lambda_i/(\mu_i - \lambda_i)$.

2.1 Demand allocation problem

A group S of stations may redistribute their arrival rates λ_i. In the sequel, we may refer to 'arrival rates' as 'demands', and use these terms interchangeably. Given an allocation of these arrival rates, the performance measure of the group is the total expected waiting time of its stations. The goal is to minimize this waiting time. The optimal performance is given by the following convex optimization problem, where the decision variables l_i, $i \in S$, are the newly allocated arrival rates.

$$\min_{l_i, i\in S} \sum_{i\in S} \frac{l_i}{\mu_i - l_i} \tag{1}$$

$$\text{s.t.} \sum_{i\in S} l_i = \sum_{i\in S} \lambda_i \tag{2}$$

$$0 \leq l_i < \mu_i,\ i \in S. \tag{3}$$

The objective value is the total expected waiting time of the customers related to the stations in the group S. Restriction (2) describes that the total arrival rate should be relocated, while restriction (3) provides the proper bounds on the arrival rate l_i allocated to station i. Denote the optimal solution by $\{l^*_{i,S}\}$ to stress the dependence on coalition S.

As a first attempt to solve this problem, consider this optimization problem without the bounds (3). Using a Lagrange multiplier α'_S this unconstrained problem may be written as follows.

$$\min_{l'_i, i \in S} \sum_{i \in S} \frac{l'_i}{\mu_i - l'_i} - \alpha'_S \left(\sum_{i \in S} l'_i - \sum_{i \in S} \lambda_i \right), \tag{4}$$

with optimal solution $\{l'_{i,S}\}$. The first order conditions for a minimum are

$$\frac{\mu_i}{(\mu_i - l'_{i,S})^2} - \alpha'_S = 0, \text{ and} \tag{5}$$

$$\sum_{i \in S} l'_{i,S} - \sum_{i \in S} \lambda_i = 0. \tag{6}$$

Equation (5) implies $\mu_i / \alpha'_S = (\mu_i - l'_{i,S})^2$. Because $l'_{i,S} < \mu_i$ (3), this results in $l'_{i,S} = \mu_i - \sqrt{\mu_i / \alpha'_S}$. Substituting this in (6) results in

$$l'_{i,S} = \mu_i - \frac{\sqrt{\mu_i}}{\sum_{j \in S} \sqrt{\mu_j}} \sum_{j \in S} (\mu_j - \lambda_j). \tag{7}$$

Hence, in this allocation station $i \in S$ receives an arrival rate $l'_{i,S}$ that starts from the stations service rate, or capacity, and subtracts from this a fraction of the overcapacity that is proportional in the square root of the capacity.

However, this solution may violate (3) by becoming negative as the example below shows.

Example 1. *Consider a Jackson network with two stations, $N = \{1, 2\}$, and parameters $\mu_1 = 16$, $\lambda_1 = 2$, $\mu_2 = 4$, $\lambda_2 = 3$. The stations minimize their total expected waiting time:*

$$\begin{aligned} \min_{l_i, i \in N} \; & \frac{l_1}{16 - l_1} + \frac{l_2}{4 - l_2} \\ s.t. \; & l_1 + l_2 = 5 \\ & 0 \leq l_1 < 16, \\ & 0 \leq l_2 < 4. \end{aligned}$$

The solution (7) to the optimization problem without lower bounds on l_1 and l_2 is $l'_{1,N} = 6$, $l'_{2,N} = -1$ with total expected waiting time 2/5. However, this solution is infeasible for station 2! Thus, the restriction $l_2 \geq 0$ in the original optimization problem will be binding. The optimal solution is $l^*_{2,N} = 0$, $l^*_{1,N} = 5$ with total expected waiting time $5/11$. *All customers go to the fastest station, station 1.*

If the unconstrained allocation problem (4) results in a solution with a negative demand allocation $l'_{i,N} < 0$ for some $i \in N$, then there is no closed-form solution to the constrained optimization problem (1)-(3). Below we present an algorithm that solves this optimization problem. This algorithm is inspired by the COOP algorithm of [3]. Given a coalition $S =$

$\{1,\ldots,s\}$ of players, the algorithm determines the optimal demand allocation $l^*_{i,S}$, $i \in S$, for this coalition.

Algorithm 1 Optimal demand allocation for $S = \{1,\ldots,s\}$.

```
Require: Queues in decreasing order of their average service rates ($\mu_1 \geq \mu_2 \geq \ldots \geq \mu_s$)
Ensure: Demand allocation $l_i$ for $i \in S$
$\Delta \leftarrow \sum_{j=1}^{s} \lambda_j$
$b \leftarrow (\sum_{j=1}^{s} \mu_j - \Delta)/\sum_{j=1}^{s} \sqrt{\mu_j}$
while $b \geq \sqrt{\mu_s}$ do
    $l_s \leftarrow 0$
    $b \leftarrow (\sum_{j=1}^{s-1} \mu_j - \Delta)/\sum_{j=1}^{s-1} \sqrt{\mu_j}$
    $s \leftarrow s - 1$
end while
for $i = 1,\ldots,s$ do
    $l_i \leftarrow \mu_i - \sqrt{\mu_i} b$
end for
```

Thus in the resulting allocation the 'slow' stations $i = s_S + 1,\ldots,s$ will have no customers ($l_i = 0$), while the 'fast' stations $i = 1,\ldots,s_S$ share the total arrival rate. The switch-over value s_S is such that $i \leq s_S$ is equivalent to $\sqrt{\mu_i} > b_S$ where

$$b_S = \frac{\sum_{j=1}^{s_S} \mu_j - \sum_{j \in S} \lambda_j}{\sum_{j=1}^{s_S} \sqrt{\mu_j}} > 0$$

is the final value of b in the algorithm. The result of the algorithm is the optimal solution to the optimization problem (1)-(3).

Theorem 1. *The demand allocation* $(l_1,\ldots,l_s)$ *resulting from Algorithm 1 is the optimal solution to* (1)-(3)*.*

Proof. The objective function (1) and the restrictions (3) are continuously differentiable and convex, and the restrictions (2) are linear. Therefore, the Karush-Kuhn-Tucker (KKT) conditions are necessary and sufficient to find an optimum. Using Lagrange multipliers α, β_i, and γ_i, $i \in S$, the minimization problem may be written as

$$\min_{l} \sum_{i \in S} \frac{l_i}{\mu_i - l_i} - \alpha(\sum_{i \in S} l_i - \sum_{i \in S} \lambda_i) - \sum_{i \in S} \beta_i l_i + \sum_{i \in S} \gamma_i(\mu_i - l_i).$$

The KKT conditions are

$$\frac{\mu_i}{(\mu_i - l_i)^2} - \alpha - \beta_i - \gamma_i = 0,\ i \in S, \tag{8}$$

$$\sum_{i \in S} l_i - \sum_{i \in S} \lambda_i = 0, \tag{9}$$

$$0 \le l_i < \mu_i,\ i \in S \tag{10}$$

$$\beta_i l_i = 0,\ i \in S, \tag{11}$$

$$\gamma_i(\mu_i - l_i) = 0,\ i \in S, \tag{12}$$

$$\beta_i \ge 0,\ \gamma_i \ge 0,\ i \in S. \tag{13}$$

Because $l_i < \mu_i$, by (12) $\gamma_i = 0$ for all $i \in S$. Then (8) reduces to

$$\frac{\mu_i}{(\mu_i - l_i)^2} - \alpha - \beta_i = 0,\ i \in S. \tag{14}$$

Consider two cases. First, if $l_i = 0$ then by (14) $\beta_i = \frac{1}{\mu_i} - \alpha$. This is nonnegative by (13), from which we conclude $\sqrt{\mu_i} \le 1/\sqrt{\alpha}$. This shows that stations with small service capacities will be inactive. This is exactly how the algorithm works. It orders the stations according to decreasing capacity, and it iteratively determines whether a station is fast enough.

Second, if $l_i > 0$ then by (11) $\beta_i = 0$. Substitute this in (14) to obtain $\alpha = \mu_i/(\mu_i - l_i)^2$, and (using $l_i < \mu_i$)

$$l_i = \mu_i - \sqrt{\mu_i}\frac{1}{\sqrt{\alpha}}. \tag{15}$$

Hence, $l_i > 0$ if and only if $\sqrt{\mu_i} > 1/\sqrt{\alpha}$. Again we see that demand is assigned to stations that have large enough service capacities. Let S_+ denote the set of all players in S with a positive demand allocation, $S_+ = \{i \in S : l_i > 0\}$. According to (9)

$$\sum_{i \in S} \lambda_i = \sum_{i \in S} l_i = \sum_{i \in S_+} (\mu_i - \sqrt{\mu_i}\frac{1}{\sqrt{\alpha}}).$$

Rewriting results in

$$\frac{1}{\sqrt{\alpha}} = \frac{\sum_{i \in S_+} \mu_i - \sum_{i \in S} \lambda_i}{\sum_{i \in S_+} \sqrt{\mu_i}} = b_S. \tag{16}$$

Substitution in (15) results in $l_i = \mu_i - \sqrt{\mu_i} b_S$. This is exactly what the algorithm does. We conclude that the algorithm returns the optimal solution. □

2.2 The arrival rate game

For the analysis of this cooperative situation, we define a related cooperative game. Given a Jackson network $(N, \{\mu_i\}, \{\lambda_i\})$, define the corresponding *arrival rate game* (N, c) with N the set of stations (players). A coalition S is a nonempty subset of N. The cost $c(S)$ of coalition S is given by its optimal performance in (1)-(3).

A cooperative game is subadditive if $c(S) + c(T) \geq c(S \cup T)$. A cost allocation x is a vector that assigns the cost x_i to player i. Such an allocation is efficient if $\sum_{i \in N} x_i = c(N)$. An allocation x is individual rational if $x_i \leq c(\{i\})$ for all $i \in N$, and it is coalitional rational if $\sum_{i \in S} x_i \leq c(S)$ for any coalition S. The core $C(c)$ of the game is the set of efficient and coalitional rational cost allocations,

$$C(c) = \{x \in \mathbb{R}^N | \sum_{i \in N} x_i = c(N);\ \sum_{i \in S} x_i \leq c(S),\ S \subseteq N\}.$$

An allocation in the core is also called a stable allocation, since no coalition has an incentive to deviate.

A collection $(w_S)_{S \subset N}$ of weights in [0,1] is a balanced collection of weights if for every player $i \in N$ the total weight of all the coalitions to which the player belongs equals 1, $\sum_{S \subset N, i \in S} w_S = 1$. A cost game (N, c) is balanced if $\sum_{S \subset N} w_S c(S) \geq c(N)$ for every balanced collection of weights. The game is called totally balanced if for any coalition S the subgame $(S, c_{|S})$ is balanced, where $c_{|S}$ denotes the restriction of the function c to players in S. Bondareva [2] and Shapley [6] independently proved that the core $C(c)$ of a cost game (N, c) is nonempty if and only if (N, c) is a balanced cost game.

Lemma 2. *Any arrival rate game is subadditive.*

Proof. Let (N, c) be an arrival rate game. The combined optimal solutions for S,$T \subseteq N$, $S \cap T = \emptyset$, are a feasible solution for $S \cup T$. The result immediately follows. □

If $l'_{i,S} \geq 0$ for all $i \in S$, then the minimal costs for coalition S are

$$c_S = \frac{(\sum_{j \in S} \sqrt{\mu_j})^2}{\sum_{j \in S}(\mu_j - \lambda_j)} - |S|,$$

where $|S|$ denotes the cardinality of S. Else, if $l'_{i,S} < 0$ for some $i \in S$ then the costs are larger than this amount, $c(S) > c_S$. For one-person coalitions, $|S| = 1$, $c(S) = c_S$.

2.3 Special case

In this subsection we consider situations in which the unconstrained allocation problem does not yield the optimal solution of the constrained allocation problem. That is, $l'_{i,S} < 0$ for some $i \in S$. We show that a slight modification of the unconstrained allocation problem yields the optimal solution.

Example 2. *Consider the situation in Example 1. The arrival rate game* (N, c) *has* $N = \{1, 2\}$, $c(\{1\}) = 1/7$, $c(\{2\}) = 3$ *and* $c(N) = 5/11$. *Because* $l^*_{2,N} = 0$, *we add this extra constraint to* (4) *and obtain*

$$\min_{l_1, l_2, \alpha, \beta_2} \frac{l_1}{16 - l_1} + \frac{l_2}{4 - l_2} - \alpha(l_1 + l_2 - 5) - \beta_2 l_2.$$

Taking derivatives to respectively l_1, l_2, α *and* β_2 *results in four equations.*

$$\frac{16}{(16 - l_1)^2} - \alpha = 0 \tag{17}$$

$$\frac{4}{(4 - l_2)^2} - \alpha - \beta_2 = 0 \tag{18}$$

$$l_1 + l_2 - 5 = 0 \tag{19}$$

$$l_2 = 0 \tag{20}$$

Combining (19) *and* (20) *gives* $l^*_{1,N} = 5$, $l^*_{2,N} = 0$. *Inserting this in* (17) *gives* $\alpha = 16/121$. *Finally, from* (18) *we derive* $\beta_2 = 57/484$. *Now consider the following allocation of costs. For station 1 consider*

$$y_1 = \frac{l^*_{1,N}}{\mu_1 - l^*_{1,N}} - \alpha(l^*_{1,N} - \lambda_1) = \frac{7}{121},$$

and for station 2

$$y_2 = \frac{l^*_{2,N}}{\mu_2 - l^*_{2,N}} - \alpha(l^*_{2,N} - \lambda_2) - \beta_2 l^*_{2,N} = \alpha\lambda_2 = \frac{48}{121}.$$

The cost allocation (y_1, y_2) *belongs to the core of the game because* $y_1 < c(\{1\})$, $y_2 < c(\{2\})$ *and* $y_1 + y_2 = c(N)$. *Hence, the core* (N, c) *is nonempty. Notice that the cost allocation* y *differs from the Shapley value and the nucleolus, which equal* $(-\frac{185}{154}, \frac{255}{154})$ *in this game.*

The cost allocation y in the example above will provide the basis of the cost allocation used in the remainder of the paper.

3 Main results

In this section we analyse the core of the arrival rate game to investigate whether it is beneficial for the stations to cooperate. We show that arrival rate games are totally balanced, and we present a cost allocation that always belongs to the core of the game.

We start by showing that the core of an arrival rate game and of all its subgames is nonempty.

Theorem 3. *Let $(N, \{\mu_i\}, \{\lambda_i\})$ be a Jackson network and (N, c) the corresponding arrival rate game. This arrival rate game is totally balanced.*

Proof. Since any subgame is another arrival rate game, it is sufficient to show that the game (N, c) is balanced.

Let $(w_S)_{S\subset N}$ be a balanced collection of weights. We show that $\sum_{S\subset N} w_S c(S) \geq c(N)$. Define $l_i^* = \sum_{S\subset N, i\in S} w_S l_{i,S}^*$, which is a convex combination of the optimal demand allocations $l_{i,S}^*$ for player $i \in N$. Then

$$\sum_{i\in N} l_i^* = \sum_{i\in N} \sum_{S\subset N, i\in S} w_S l_{i,S}^* = \sum_{S\subset N} w_S \sum_{i\in S} l_{i,S}^* = \sum_{S\subset N} w_S \sum_{i\in S} \lambda_i = \sum_{i\in N} \lambda_i \sum_{S\subset N, i\in S} w_S = \sum_{i\in N} \lambda_i.$$

Using similar reasoning, $0 \leq l_i^* < \mu_i$. Therefore, the demand allocation $\{l_i^*\}$ is a feasible solution of the allocation problem for coalition N. Finally,

$$c(N) \leq \sum_{i\in N} \frac{l_i^*}{\mu_i - l_i^*} \leq \sum_{i\in N} \sum_{S\subset N, i\in S} w_S \frac{l_{i,S}^*}{\mu_i - l_{i,S}^*} = \sum_{S\subset N} w_S \sum_{i\in S} \frac{l_{i,S}^*}{\mu_i - l_{i,S}^*} = \sum_{S\subset N} w_S c(S),$$

where the first inequality follows from $\{l_i^*\}$ being a feasible solution, and the second inequality is due to the convexity of the function $\frac{l_i}{\mu_i - l_i}$ in l_i. We conclude that the game is balanced. □

This total balancedness implies that cooperation is beneficial for the stations in the Jackson network. Notice that the proof above is similar to the proof of the theorem of Shapley and Shubik [7], when they show that any market game is balanced. The main difference is due to our extra bounds on the demand allocation (3).

Based on our analysis in Section 2.3 and Theorem 1 we propose the following cost allocation. Player i pays

$$x_i = \frac{l_{i,N}^*}{\mu_i - l_{i,N}^*} - \frac{1}{b_N^2}(l_{i,N}^* - \lambda_i), \; i \in N, \tag{21}$$

with $l_{i,N}^*$ the optimal demand allocation for player i in the allocation problem (1)-(3) of coalition N. Thus each player has to pay its individual cost in the optimal solution, it enjoys a cost reduction if it serves more demand than its original arrival rate, and otherwise it pays extra. This allocation belongs to the core of the arrival rate game.

Theorem 4. *Let $(N, \{\mu_i\}, \{\lambda_i\})$ describe a Jackson network with corresponding arrival rate game (N, c). The cost allocation x in (21) belongs to the core of the game.*

Proof. Efficiency of x follows immediately from the definition,

$$\sum_{i \in N} x_i = \sum_{i \in N} \left(\frac{l^*_{i,N}}{\mu_i - l^*_{i,N}} - \frac{1}{b^2_N}(l^*_{i,N} - \lambda_i) \right) = \sum_{i \in N} \frac{l^*_{i,N}}{\mu_i - l^*_{i,N}} = c(N).$$

For coalitional rationality, let α_N, $\beta_{i,N}$ be the values of α and β_i that satisfy the KKT conditions for coalition N. Consider the function

$$f_i(l_i) = \frac{l_i}{\mu_i - l_i} - \alpha_N(l_i - \lambda_i) - \beta_{i,N} l_i.$$

Taking the derivative and equating it to zero, results in

$$\frac{\mu_i}{(\mu_i - l_i)^2} - \alpha_N - \beta_{i,N} = 0.$$

This resembles the KKT condition (14). Hence, the function f_i is minimized in $l_i = l^*_{i,N}$. Note that (16) for $S = N$ and (11) imply $x_i = f_i(l^*_{i,N})$.

Let S be a coalition with optimal demand allocations $\{l^*_{i,S}\}$. Then $x_i = f_i(l^*_{i,N}) \leq f_i(l^*_{i,S})$ because $l^*_{i,N}$ is a minimizer of f_i. Summing over the players in S gives

$$\sum_{i \in S} x_i \leq \sum_{i \in S} f_i(l^*_{i,S}) = c(S) - \sum_{i \in S} \beta_{i,N} l^*_{i,S} \leq c(S),$$

where the latter inequality is due to (3) and (13). Hence, x satisfies coalitional rationality. We conclude that $x \in C(c)$. □

The tailor-made cost allocation x is a stable cost allocation in the arrival rate game. This Theorem also implies that arrival rate games are totally balanced, thus it is a second way of proving the result of Theorem 3.

If the demand allocations $\{l'_{i,N}\}$ are nonnegative, we obtain the following alternative formulation of the cost allocation x.

Corollary 5. *Let $(N, \{\mu_i\}, \{\lambda_i\})$ be a Jackson network with arrival rate game (N, c). If $l'_{i,N} \geq 0$ for all $i \in N$, then*

$$x_i = \frac{l'_{i,N}}{\mu_i - l'_{i,N}} - \alpha'_N(l'_{i,N} - \lambda_i)$$

is an alternative formulation of the cost allocation x.

References

[1] S.M. Alavi, C. Zhou. Resource allocation scheme for orthogonal frequency division multiple access networks based on cooperative game theory. *International Journal of Communication Systems*, 27: 1105 – 1125, 2014.

[2] O.N Bondareva. Some applications of linear programming methods to the theory of games. (In Russian) *Problemy Kibernet* 10: 119 – 139, 1963.

[3] D. Grosu, A.T. Chronopoulos, M.Y. Leung. Cooperative load balancing in distributed systems. *Concurrency and Computation: Practice and Experience*, 20: 1953 – 1976, 2008.

[4] F. Karsten, M. Slikker, G.-J. van Houtum. Resource pooling and cost allocation among independent service providers. *Operations Research*, 63: 476 – 488, 2015.

[5] U. Özen, M.I. Reiman, Q. Wang. On the core of cooperative queueing games. *Operations Research Letters*, 39: 385 – 389, 2011.

[6] L.S. Shapley. On balanced sets and cores. *Naval Research Logistics Quarterly*, 14: 453 – 460, 1967.

[7] L.S. Shapley, M. Shubik. On market games. *Journal of Economic Theory*, 1: 9 – 25, 1969.

[8] A.N. Tantawi, D. Towsley. Optimal static load balancing in distributed computer systems. *Journal of the Association for Computing Machinery*, 32 (2): 445 – 465, 1985.

[9] J. Timmer, W.R.W. Scheinhardt. How to share the cost of cooperating queues in a tandem network? Conference proceedings of the 22nd International Teletraffic Congress 2010, *IEEE Xplore*. DOI: 10.1109/ITC.2010.5608712.

[10] J. Timmer, W.R.W. Scheinhardt. Cost sharing of cooperating queues in a Jackson network. *Queueing Systems*, 75 (1): 1 – 17, 2013.