# Modèles de files d'attente pour l'analyse des stratégies de collaboration dans les systèmes de services

Jing Peng

NNT : 2016SACLC089

# THESE DE DOCTORAT
DE
L'UNIVERSITE PARIS-SACLAY
PREPAREE A
"CENTRALESUPELEC"

ECOLE DOCTORALE N°573
Interfaces : approches interdisciplinaires / fondements, applications et innovation

Sciences et technologies industrielles

Par

**Mlle. Jing PENG**

## Queueing approaches for the analysis of collaboration strategies in service systems

**Thèse présentée et soutenue à Châtenay-Malabry, le « 19/12/2016 »:**

**Composition du Jury :**

| | | |
|---|---|---|
| M. Marc AIGUIER | Professeur, CentraleSupélec | Président |
| M. Jean-Philippe GAYON | Maître de conférences, HDR, l'INPG | Rapporteur |
| M. Yacine REKIK | Professeur, EMLYON | Rapporteur |
| M. Fabrice CHAUVET | HDR, EDF Recherche & Développement | Examinateur |
| M. Yves DALLERY | Professeur, CentraleSupélec | Directeur de thèse |
| M. Zied JEMAI | Maitre de conférences, HDR, OASIS-ENIT | Co-encadrant de thèse |
| M. Oualid JOUINI | Professeur, CentraleSupélec | Co-encadrant de thèse |

**Université Paris-Saclay**
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France

# Remerciement

A l'issue de la rédaction de cette recherche, je n'aurais jamais pu réaliser ce travail doctoral sans le soutien d'un grand nombre de personnes: l'équipe encadrante de thèse, mes amis, mes collègues et mes parents.

Premièrement, je tiens à remercier mon directeur de thèse, monsieur *Yves DALLERY*, pour la confiance qu'il m'a accordée en acceptant de diriger ce travail doctoral. Je voudrais remercier sincèrement mes encadrants, monsieur *Oualid JOUINI* et monsieur *Zied JEMAI*, pour leurs multiples conseils et toutes les heures quils ont consacrées à encadrer cette recherche.

Mes remerciements vont également à monsieur *Marc AIGUIER*, monsieur *Jean-Philippe GAYON*, monsieur *Yacine REKIK*, et monsieur *Fabrice CHAUVET* pour avoir accepté de participer à ce jury de thèse, pour avoir bien voulu porter à mon travail, pour l'occasion de discuter du résultat de mes recherches et les recherches possibles en perspective avec eux et pour tous les conseils intéressants sur mon travail.

Je voudrai remercier le *China Scholarship Council* qui a financé cette thèse.

Mes remerciements vont aussi à mes amis et mes collègues du laboratoire qui avec cette question récurrente $\ll$ quand est-ce que tu la soutiens cette thèse? $\gg$, m'ont permis de ne jamais dévier de mon objectif final. Merci à *Delphine* et *Corinne* pour tous les aides pendant le processus de soutenance. Je remercie également tous les participants de ma

soutenance de thèse pour leurs présences à la date spécifique avant Noël.

Ma reconnaissance va à ceux qui ont plus particulièrement assuré le soutien affectif de ce travail doctoral : ma famille. Ma mère m'encourage énormément quand j'ai rencontré la crise la plus importante pendant la thèse. Je souhaite enfin remercier particulièrement mes amis qui m'ont accompagné et m'ont aidé beaucoup pendant ma vie doctorale, *Wenjing*, *Shanshan*, *Huan*, *Xue* et *Shouyu*, malgré certaines ont déjà rentrés en Chine.

***Jing PENG***

# Contents

# Chapter 1

# Introduction

This chapter provides a general introduction to this Ph.D. thesis. It is divided into three main sections. First, we present the background and the motivations of our work. Second, we highlight our objective and contributions. Finally, we describe the structure of the manuscript.

## 1.1. Background & motivation

Many of our daily activities depend on services and service providers, from the e-mail we check in the early morning to the public transportation service we take to our working place, from the restaurant we eat in at noon to the package we receive during the day. Services are everywhere in our life, including finance (banking, stocks), health (personal physician, hospital), communication (e-mail, 4G network), public services (electricity, police), etc. [Daskin, 2010]

In past twenty years, the service sector has emerged as the primary sector in the world economy (Figure 1.1), especially in developed countries. Based on the report from the office of the United States Trade Representative, for instance, four out of five jobs in the U.S. are provided by service industry [USTR, 2014]. Furthermore, service sector accounted for 78.76% gross domestic product (GDP) in France 2015 according to statistical data from the World Bank group.

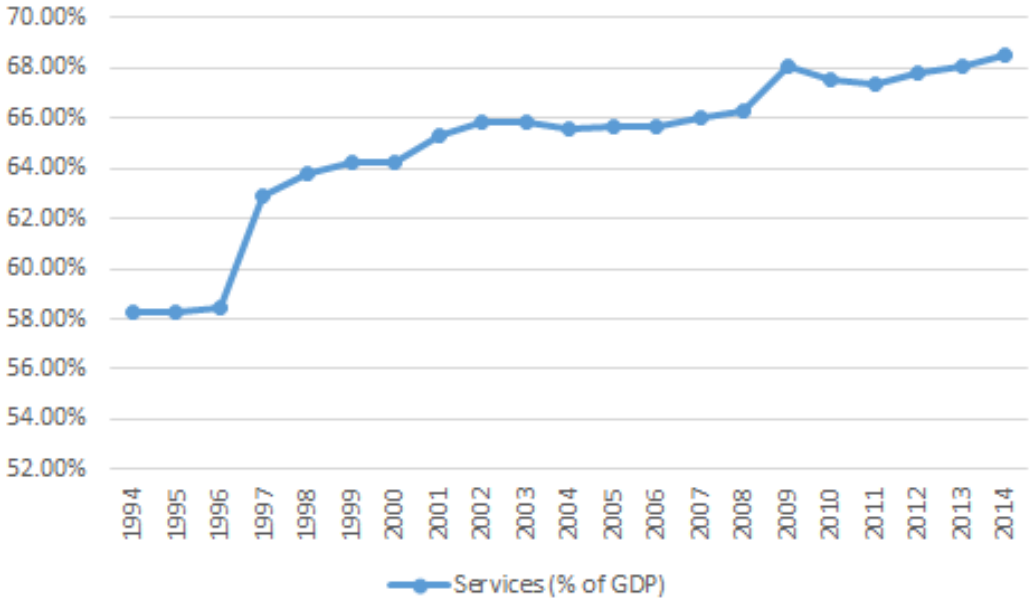



**Figure 1.1:** *Growth of the service sector in world GDP*

In the context of economic globalization, competition and cooperation in service industries have become more and more popular: price competition among fast food

restaurant chains, combination operation of telecommunication companies, collaborative after-sales and maintenance services in electronic manufacturing industry, just to name a few. In this thesis, we study collaborative strategies in homogeneous service systems. We focus in particular on resource pooling strategies. Our approach consists of using queueing modeling for service systems and game theory for the analysis of interactions between service providers. In what follows, we briefly discuss collaboration strategies and resource pooling.

**Collaboration strategies in services**

In order to improve the system performance or reduce expenses, there are several basic cooperative methods: queueing cooperation, e.g., scheduling among simultaneous arrival agencies or rerouting among different servers [Katta and Sethuraman, 2006, Kayi and Ramaekers, 2010]; service pooling, e.g., service rate pooling or staffing allocation [Guo et al., 2013]; cross-training [Tekin et al., 2014]; collaboration with third-party service providers, e.g., service outsourcing [Aksin et al., 2008], etc. It is sometimes useful to combine these methods to form a more profitable collaborative structure [Anily and Haviv, 2014].

Cooperation methods can be classified broadly into three typical forms (Figure 1.2): vertical form, the collaboration between customers and servers, e.g. phone packages signed with telecommunication companies, fitness cards brought from gyms; horizontal form, the collaboration among homogeneous servers, e.g. after-sale services of electronic products of different brands; and external form, the collaboration with another party out of the service systems, e.g., customer services outsourcing abroad.

Among majority efficiencies brought by service collaborations, cost reduction is the most marked driver for service providers. The service capacity cost and the waiting cost in the queue/system are widely used in the literature [Anily and Haviv, 2010, Özen et al., 2011, Karsten et al., 2015b, Yu et al., 2015].

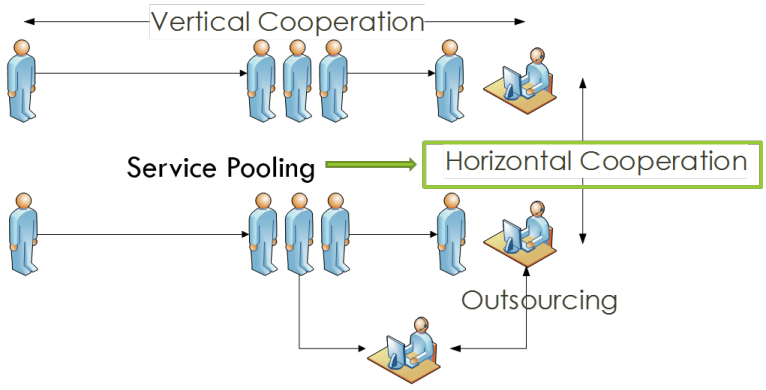



**Figure 1.2:** *Service cooperation classification*

### Resource pooling

From first study in [Stidham, 1970], pooling for queueing systems has been widely investigated in the literature on the design of service systems. It is well known that the service capacity pooling naturally leads to economies of scale in stochastic flows in operation management studies [Smith and Whitt, 1981, Bell and Williams, 2005]. This operational efficiency improvement occurs in the disappearance of idle service resources in the presence of congestion in queues. It is both valid within some departments of an economic entity, e.g., reservation pooling in a restaurant [Thompson and Kwortnik, 2008], or among multiple independent entities [González and Herrero, 2004, Garcia-Sanz et al., 2008, Anily and Haviv, 2010, Kayi and Ramaekers, 2010, Tekin et al., 2014, Anily and Haviv, 2014].

Applications in practice for service pooling among homogeneous service providers are numerous. For instance, different departments in a hospital could share a common operating theatre and afford the joint expenses. Different hospital departments could also share a joint service capacity in terms of beds in a common ward, which would alleviate congestion. Another example is in the context of after-sales for new categories of electronic products. Such products are likely to have low after-sales demand rates for each retailer individually. The retailers could therefore provide together a joint after-sales service to reduce service start-up costs and also improve service quality. For aviation services, the joint check-in service for different airline companies is an additional

example for service pooling applications.

Prior to services, resource pooling in supply chains has already attracted a lot of attention. The first contribution to gains splitting is considered in a multistore economic order quantity with safety stock in [Gerchak and Gupta, 1991]. Later, [Hartman and Dror, 1996] and [Özen et al., 2008] extend this problem in a cooperative game environment. The cooperation costs sharing issue in the multi-retailer newsvendor problem is first considered by [Hartman et al., 2000] and [Müller et al., 2002]. Relative problem are also fruitfully studied in the joint replenishment problem [Meca et al., 2004, Anily and Haviv, 2007, Zhang, 2009, Elomri et al., 2012] and the economic lot-sizing model [Van den Heuvel et al., 2007, Guardiola et al., 2009].

There are similarities between service and manufacturing operations, which are both concerned with the efficiency, effectiveness, quality problems, and motivated by the cost reduction. In contrast to the research in the manufacturing industry, the relative research in service industry could not meet requirements of its enormous economic share. Services are mainly characterized by complex operations and a high impact of human factors. In this thesis, we account for these two aspects through the analysis of the impact of service duration variability and customer abandonment, respectively. We study the problem where independent service providers could be subject to cooperate with each other. We consider the resource pooling strategy in different service systems and provide corresponding pooling strategies using cooperative game theory.

## 1.2. Objective & contributions

The objective of this thesis is to study the impact of the features of service variability and customer abandonment on collaboration strategies. Motivated by cost reduction, we tackle the resource pooling problem between independent service providers. We use a queueing approach for the modeling of these features. More concretely, we address the two following questions: 1) which coalition strategy should be used? and 2) which allocation rule should be selected in order to maintain the stability of the coalition? We

use cooperative game theory, which provides interesting concepts to analyze profitable coalition structures and solve the cost-sharing problem among the participants.

The main contributions of this thesis can be summarized as follows.

First, we study the cost-sharing problem among independent service providers in a service capacity pooling system with general service times. The effective improvement is achieved by reducing the resource idleness in case of congestion. We model both the service provider and the cooperative coalition as single server queues with general service times. For the two situations of pooling with a fixed service capacity and pooling with the optimized service capacity, we define the corresponding cooperative games and analyze the core allocations. For the fixed capacity case, we prove that the core is non-empty. The characteristic function is neither concave nor monotone in the aforementioned game. However, we prove that the service pooling game with the optimized service capacity is concave. For this concave game, we find two stable allocation rules and illustrate a combined cost allocation strategy.

Second, we consider a group of homogeneous and independent single server service providers with impatience, where a customer quits the system without service whenever her waiting time in the queue exceeds his patience time threshold. The advantage of collaboration in the service systems accounting customer abandonment, is not only the sharing of instant idle resources but also the reducing of abandoned customers. Under Markovian assumptions for inter-arrival, service and patience times, we define a cooperative game with transferable utility and a fixed service capacity for each coalition. We prove that the grand coalition is the most profitable coalition and that the game has a non-empty core. We then examine the impact of abandonment on the stability of Shapley value. Furthermore, we prove the concavity of the waiting queue length with respect to the abandonment rate, and give a condition under which the Shapley value is situated in the core. We also study the cost-sharing problem of the relative cooperative game with the optimized service capacity, and prove that the proportional allocation rule based on customer arrival rates gives a dynamic stable allocation to all relative sub-games.

In the previous studies, we use the 'super-server' assumptions. The main reason

for this assumption is that dealing with multi-server queues with general service times and customer abandonment is very hard. To assess the quality of this assumptions, we address the service pooling problem in the multi-server pooling setting. Although it is intuitive to expect efficiency improvements in the pooled multi-server system, it is not obvious to conclude that all members will benefit from pooling as it is the case for the 'super-server'. We compare between two pooling settings from a coalition perspective. We numerically evaluate the effects of service variability and customer abandonment on the two corresponding games.

## 1.3. Thesis structure

The remaining part of this Ph.D dissertation contains four chapters.

In Chapter 2, we study the cost-sharing problem in service systems with general service times. The paper version of this chapter is under the second round review in *Naval Research Logistics* [Peng et al., a]. Secondly, we analyze cooperative strategies in the presence of customer abandonment in Chapter 3. The paper version of this chapter is submitted to *IIE Transactions* [Peng et al., b]. In Chapter 4, we compare between 'super-server' and multi-server modeling in terms of stability and cost allocations. The paper version of this chapter is a working paper to be submitted for publication. Finally, Chapter 5 is devoted to general conclusions and perspectives. Figure 1.3 provides a general overview of the dissertation content.

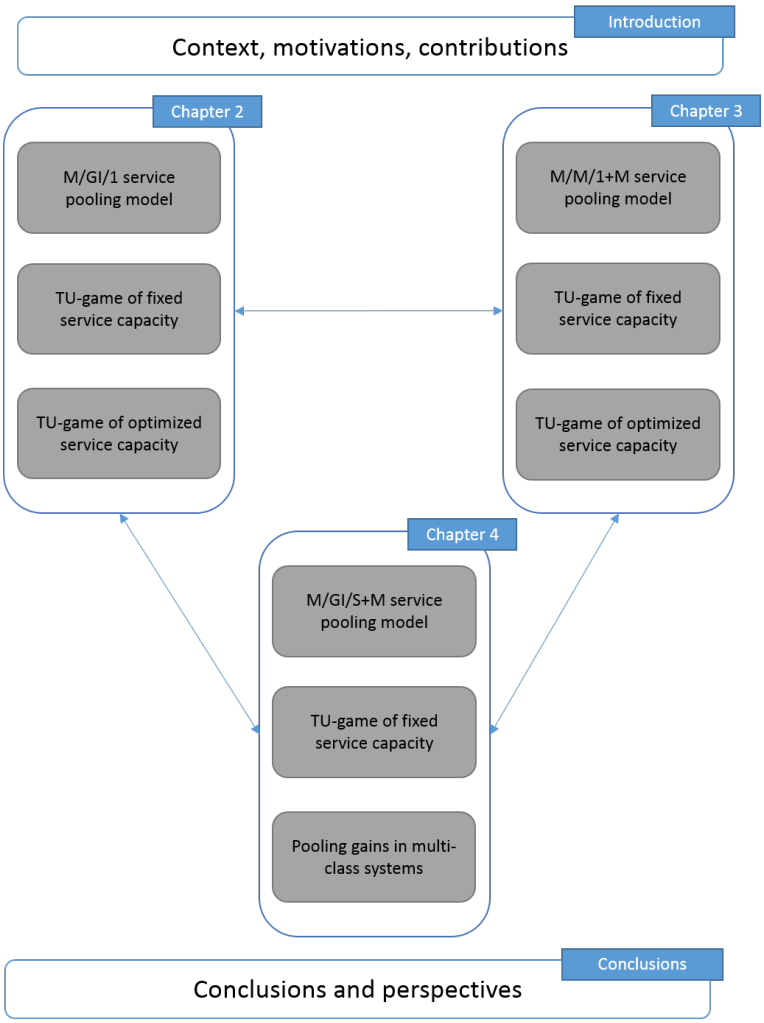



**Figure 1.3:** *Dissertation structure*

# Chapter 2

# Cooperation in Service Systems with General Service Times

In this chapter, we study the cost-sharing problem among independent service providers in a service capacity pooling system with general service times. The effective improvement is achieved by reducing the resource idleness in case of congestion. We model both the service provider and the cooperative coalition as single server queues with general service times, and attempt to answer the following questions: 1) which coalition strategy should be used; and 2) which allocation rule should be selected in order to maintain the stability of the coalition?

For both situations, (a) pooling with a fixed service capacity and (b) pooling with the optimized service capacity, we define the corresponding cooperative game and analyze the core allocations. For the fixed capacity case, we prove that the core is non-empty. The characteristic function is not concave and monotone in

the aforementioned game. However, we prove that the service pooling game with the optimized service capacity is concave. For this concave game, as it is widely admitted, the Shapley value provides a core allocation rule. We prove that the proportional allocation rule based on the individual customer arrival rates, is also located in the core. Moreover, we show that the allocation scheme evolved is a Population Monotonic Allocation Scheme (PMAS) for this game. We finally analytically compare between the performance of the Shapley value and the proportional rule in terms of the individual profits from the cooperation, and illustrate a combined allocation strategy.

## 2.1. Introduction

The obvious profit of service pooling is the congestion mitigation in the whole system, owing to the reduction of idleness with the presence of waiting customers. The pooling advantage for the entire alliance is apparent, but the collective interests cannot be the incentive for each individual service provider to join the coalition. It is therefore important to address the following questions: which service providers should cooperate; and how to share the pooling cost among the participants to keep each individual and subset staying in the coalition?

Motivated by real-life applications, we consider in this chapter a set of independent single server service providers, each of which faces its own incoming stream of customers. Customer inter-arrival and service times are assumed to be random and independently distributed. We suppose that every incoming stream is strictly unrelated to those of other providers. This means that there is no competition in the set. Service providers could then join a profitable coalition by operating their service capacities in common. Alternatively, each provider makes his own decision independently to either join any coalition or not, based on his individual benefit. Once the coalition is formed, the most interesting problem for every unit in the entire coalition becomes a cost-sharing problem.

Cooperative game theory provides interesting concepts to look for profitable coalition structures and solve the cost-sharing problem among the participants. We assume here that the total cost is a transferable utility, e.g., money in the general case. The corresponding cooperative game with transferable utility (TU-game) is defined among a set of independent service providers, and has a characteristic function defined by the operating coalition costs. We prove that the service pooling game with fixed service capacity sharing always has a stable cost-sharing solution for the grand coalition. Stable cost-sharing means that all subsets pay less in the grand coalition than in each individual setting. We also observe that the higher is the variability of service times, the larger is the relative revenue to the cooperative coalition. Under optimized service capacity conditions, we prove that the corresponding service pooling game is concave. We consider

two stable cost allocation rules for this game: the proportional allocation rule depending on customer arrival rates and the Shapley value, and discuss their fairness using the benefit ordering property.

The rest of this chapter is structured as follows. We briefly review the relevant literature in the next section. In Section 2.3, we present the individual and collaborative modeling of service systems. In Section 2.4, we define and analyze the service pooling problem with a fixed service capacity as a TU-game. Then, we consider the optimal service rate and analyze the corresponding service pooling game in Section 2.5. We consider two stable allocation rules for this game and provide some analytical discussions of the results.

## 2.2. Literature review

Our work is related to the stream of literature dealing with the study of the benefits of resource pooling. In the early research [Stidham, 1970], the optimal design of single server systems is studied for different service cost functions. Moreover, the resource pooling as a parallel-server system is applied in the heavy transportation case in [Harrison and López, 1999]. In [Wallace and Whitt, 2004], the authors consider both the resource pooling and staffing in a particular call center application. While focusing on the profitability, [Dijk and Sluis, 2008, Jouini et al., 2008] discuss the benefit of pooled and unpooled scenarios in call centers. For an inventory application, the sensitivity of the inventory pooling benefit is investigated and evaluated by comparing several forms of capacity pooling in [Benjaafar et al., 2005].

This work is also related to the large body of literature focusing on the cooperative behavior among independent participants using cooperative game theory. [González and Herrero, 2004] is the earliest research that deals with the cost-sharing problem for an operating theatre in medical service. The authors separate the operating theatre costs as variable and fixed costs, and focus on the Shapley value of two sub-games. In [Garcia-Sanz et al., 2008], the authors extend the work in [González and Herrero, 2004] by con-

sidering preemptive priority for the customers from different individual servers, which allows to provide a more profitable pooling system. Our work is clearly related to two recent papers [Yu et al., 2015, Anily and Haviv, 2010] that consider M/M/1 modeling to study the service capacity pooling problem using cooperative game theory. [Yu et al., 2015] uses the optimal service rate that minimizes the system operating costs, and an incomplete information problem is also treated. In [Anily and Haviv, 2010], the authors choose a service rate varying with the customer incoming rate. In [Karsten et al., 2015b] and [Karsten et al., 2011], the service collaboration problems in [Yu et al., 2015] and [Anily and Haviv, 2010] are extended to multi-server settings using Erlang-C and Erlang-B queueing models, respectively. Similar to [Karsten et al., 2015b, Karsten et al., 2011], we consider here a single server modeling. Yet, the cost structure is different since we focus on the cost of the waiting time in the queue instead of that in the system. This makes the results different. For instance, the concavity of the auxiliary game provided in [Anily and Haviv, 2010] is not compatible with our setting. More importantly, we allow service time to be generally distributed. Despite its prevalence in practice, no existing studies allow for non-Markovian service times.

The cost allocation problem of service capacity pooling is a challenging subject. The objective is to find the stable cost allocations, which means that no coalition has an incentive to split off. In cooperative games, the core defined in [Gillies, 1959], presents the set of all stable cost allocations. For service pooling games, the Shapley value as an accepted fair cost allocation rule is discussed in [González and Herrero, 2004, Anily and Haviv, 2010]. Note also that the proportional allocation rule could be used as a general cost-sharing rule for this kind of problems. It depends on the structure of the game characteristic function [Garcia-Sanz et al., 2008, Yu et al., 2015, Karsten et al., 2015b, Karsten et al., 2011]. In [Anily and Haviv, 2010], the authors divide all stable allocations into two families: non-negative and negative ones, and propose an algorithm to generate all stable allocations. A corrected proportional allocation rule is given in [Yu et al., 2015] to handle the eventual incomplete information problem. For our game with the optimized service capacity, we find two stable allocation rules. Considering the

complexity and the fairness of the two rules, we discuss their use under different setting.

There are also several papers focusing on other service collaboration issues using cooperative game theory: [Katta and Sethuraman, 2006] considers a scheduling problem in a service facility under a rush hour regime. They propose two solution concepts, Random Priority and Constrained Random Priority cores, if monetary compensation are allowed. For a similar scheduling problem, the only cost allocation rule satisfying Pareto-efficiency, anonymity and strong strategy-proofness, is proved by [Kayi and Ramaekers, 2010]. In [Mishra and Rangarajan, 2007], the authors characterize the Shapley value solution for a job scheduling problem. These kind of problems are referred to as queueing problems in [Chun, 2006b, Chun, 2006a]. [Guo et al., 2013] studies the staffing problem in call centers, where a square-root safety staffing rule is selected to define the number of agents required for each coalition in the system. They show that the Shapley value is a fair staff allocation rule in the core of the square-root safety staffing game.

In addition to services, cooperation in supply chains has been already a fruitful research subject to make joint pricing and inventory decisions. In order to improve operations and reactivity to the market changes, working with outside partners is more and more important for supply chain actors over the past decades [Zhang, 2009, Dror et al., 2012, Elomri et al., 2012, Timmer et al., 2013]. There are numerous examples for cooperation cases: collaborative forest transportation [Fiestras-Janeiro et al., 2013], cooperative procurement [Drechsel and Kimms, 2010], shipping pooling of automobiles [Sherali and Lunday, 2011], lateral transshipments pooling [Satir et al., 2012], just to name a few. A interesting review of the application of both cooperative and non-cooperative game theories in supply chain is provided by [Nagarajan and Sošić, 2008].

## 2.3. Service pooling modeling with general service times

We consider a set of $n$ service providers, $N = \{1, \ldots, n\}$. Each service provider $i \in N$ is modeled as a single server queue handling a single class of customers. We assume that the waiting space is large enough, no customer would abandon after arriving at

the system, and there is no failure in service processing, i.e., no retrial is considered here. The incoming stream of customers to service provider $i \in N$ follows a Poisson process, and customers are served in the order of their arrivals, i.e., under the first come, first served (FCFS) discipline of service. Service times for a given service provider are assumed to be i.i.d. and allowed to follow a general distribution. Following the above assumptions, the individual service process is modeled as an M/GI/1 queueing system. For the total operating cost, we consider a traditional economic framework with two types of costs [Guajardo and Rönnqvist, 2015]. The first type is a linear capacity cost per unit time, which is proportional to system service capacity. This may capture the equipment's depreciation or maintenance fee, employee's salary, etc. We also assume a congestion cost incurred for each unit of time the customers spend in the queue. For each service provider $i$, we define the following parameters:

- $\lambda_i$: Mean arrival rate of customers to provider $i$;
- $T_i$: A random variable describing the service time at server $i$, with mean $1/\mu_i$ and coefficient of variation $cv_i$;
- $\rho_i = \lambda_i/\mu_i$: Server utilization for provider $i$, with $\mu_i > \lambda_i$;
- $W_{q,i}$: Customer expected waiting time in the queue for provider $i$;
- $c_{h,i}$: Service capacity cost parameter per unit time per service capacity for provider $i$;
- $c_{w,i}$: Congestion cost parameter per unit time per customer waiting in the queue at server $i$.

For service provider $i$, we denote the capacity, the congestion and the total operating costs per unit time by $C_{h,i}$, $C_{w,i}$ and $C_i$, respectively. Using the *Pollaczek-Khinchine formula* in [Pollaczeck, 1930], we may write

$$C_i = C_{h,i} + C_{w,i} = c_{h,i}\mu_i + c_{w,i}\lambda_i W_{q,i} = c_{h,i}\mu_i + c_{w,i}\lambda_i \frac{\lambda_i \mathbf{E}(T_i^2)}{2(1-\rho_i)}. \tag{2.1}$$

The service capacity pooling consists of two typical methods with pooling demand. In the first one, the service providers form a common facility with parallel-servers and one single queue, which will be considered in Chapter 4. In the second one, the servers

share their service capacities together and work as a 'super-server', i.e., a single server system with a high service capacity. As in [Yu et al., 2015] and [Anily and Haviv, 2010], we consider here the second configuration. We assume that $u$ independent service providers of any subset $\emptyset \subset U = \{1, \ldots, u\} \subseteq N$ would decide to share their capacities as a 'super-server'. Individual arrival processes are assumed to be independent. Therefore, the combined arrival process, at a rate of $\lambda_U = \sum_U \lambda_i$ follows a Poisson process. We assume that the pooling system provides same services in terms of speed and quality (compared to the case of individual providers), and service times are also i.i.d.. We denote the mean service capacity by $\mu_U$ for the pooling system $U$. Based on these assumptions, the $u$ providers act as an M/GI/1 'super-server' system (Figure 2.1).

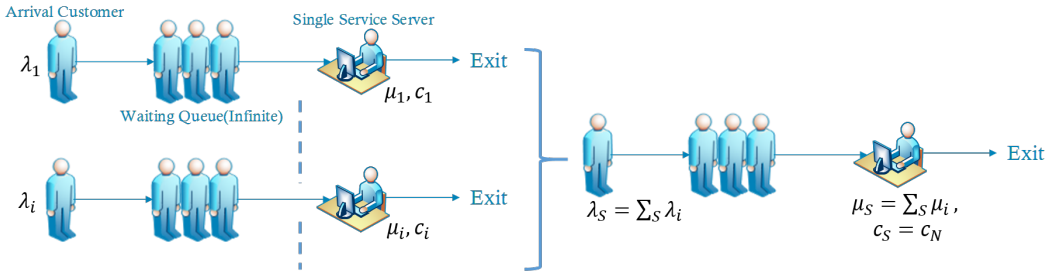



**Figure 2.1:** *Individual and coalition service systems*

The service providers, which provide the same service and have similar scales, are assumed to also have similar service capacity expenses. This could be justified by the use of the same technology and the similarity of the staff salary expenses. We then assume for simplicity that the service capacity cost parameters to be the same, $c_{h,i} = c_h$, and the coefficients of variation of service times to be identical, $cv_{U,i} = cv_U = cv_N$, for all providers in the set $N$. We only consider a single class of customers in the pooling system. The congestion parameters are assumed to be identical, $c_{w,i} = c_w$, for all customers from all service servers. The service providers are also all assumed to be risk-neutral. The total cost of the pooling system $U \subset N$, is the sum of service capacity cost and system congestion cost. For a pooling system $U$, we define the parameters as follows:

- $\lambda_U$: Mean arrival rate of customers for the pooling system $U$;
- $T_U$: A random variable, service time at the pooling server $U$, with mean $1/\mu_U$ and

coefficient of variation $cv_U$;

- $\rho_U = \lambda_U/\mu_U$: Server utilization for the pooling system $U$, with $\mu_U > \lambda_U$;
- $W_{q,U}$: Customer expected waiting time in the queue for the pooling system $U$;
- $c_h$: Service capacity cost parameter per unit time per service capacity for any subset of $N$;
- $c_w$: Congestion cost parameter per unit time per customer waiting in the queue for any subset of $N$.

For a pooling system $U$, we denote the capacity, the congestion and the total operating costs per unit time by $C_{h,U}$, $C_{w,U}$ and $C_U$, respectively. The expected total cost per unit time is given by

$$C_U = C_{h,U} + C_{w,U} = c_h\mu_U + c_w\lambda_U W_{q,S} = c_h\mu_U + c_w\lambda_U \frac{\lambda_U \mathbf{E}(T_U^2)}{2(1-\rho_U)}. \tag{2.2}$$

With the above notations and definitions, we propose both the individual service model and the pooling service model using an M/GI/1 queueing modeling. Our objective is to investigate the profitability and stability of service capacity pooling. It consists of two steps: (i) confirm that the service capacity pooling is profitable for the whole set and for each provider; (ii) verify that the coalition structure could be stable under some cost allocations. Cooperative game theory provides a framework to formulate, structure and analyze the cooperative behavior of independent individuals in collaboration. In the two following sections, we construct TU-games under two different conditions: (i) each provider has a fixed individual service capacity and the capacities remain fixed in any coalition; (ii) the service capacity for any individual/pooling server could be optimized.

## 2.4. Service pooling game with fixed service capacity

In this section, we assume that the service capacity of every individual service provider is fixed. This corresponds to situations where the changing of equipments or the physical location is too expensive or almost impossible. We thus consider the pooling capacity $\mu_U$ of any subset $U$ as the sum of the service capacities of its members. Let us denote by

$C_{fix}(\cdot)$ the total cost function of a subset $U \subset N$. We have

$$C_{fix}(\emptyset) = 0, C_{fix}(U) = C_U(\lambda_U, \mu_U = \sum_{j \in U} \mu_j), \text{for any } \emptyset \subseteq U \subseteq N. \tag{2.3}$$

Consider a finite set of independent service providers $N = \{1, \ldots, n\}$ (a set of players), which is known as the grand coalition in cooperative game. Let $U$ be any subset of $N$, which is called a coalition (a subset of players). In every coalition, the players could reduce the total costs by service pooling. We assume that each service provider could only join one coalition, and the pooling cost could be redistributed among the providers with no limitation. Thus, a TU-game $(N, C_{fix})$ is completely specified by its characteristic function $C_{fix} : 2^N \to \mathbf{R}$ defined for any sub-coalition of $N$.

To simplify the presentation of the game $(N, C_{fix})$, we define the quality $f$ as $cv_N^2$ and rewrite Equations (2.1) and (2.2). We thus obtain

$$C_U = c_h \mu_U + c_w \lambda_U \frac{\lambda_U \mathbf{E}(T_U^2)}{2(1-\rho_U)} = c_h \mu_U + \frac{c_w(1+f)}{2} \frac{1}{\rho_U^{-1}(\rho_U^{-1} - 1)}, \text{for any } \emptyset \subseteq U \subseteq N. \tag{2.4}$$

### 2.4.1. Non-emptiness of the core

Note here that the service capacity cost and the system congestion cost are both single-variable functions in either the service capacity or the server utilization. We denote the expected queue length of coalition $U$ by $L_{q,U}$. Proposition 2.1 states the improvement of the service quality in terms of $L_{q,U}$ in the pooling system.

**Proposition 2.1.** The expected queue length is strictly subadditive in the pooling system with a fixed service capacity.

*Proof.* For any $\emptyset \subseteq U, T \subseteq N$ with $U \cap T = \emptyset$, we suppose that $\rho_U \leq \rho_T$. Then, the utilization of the pooling server $U \cup T$ has the property: $\rho_U \leq \rho_{U \cup T} \leq \rho_T$, and the queue length $L_{q,U} = (1+f)/[2\rho_U^{-1}(\rho_U^{-1} - 1)]$ is an increasing function in $0 < \rho_U < 1$. Thus, $L_{q,U} \leq L_{q,U \cup T} \leq L_{q,T}$. So, $L_{q,U \cup T} < L_{q,U} + L_{q,T}$. The subadditivity of the expected queue length has been proved. □

From Proposition 2.1, we state that the average number of waiting customers is always reduced in the pooling system. The proof above also shows that $L_{q,U}$ is not monotone. The joint queue length may be increased with the joining of new members, e.g., when a provider $i \notin U$ joins a coalition $U$ with $\rho_U < \rho_i$. This result is similar to that in the M/M/1 service pooling game in [Anily and Haviv, 2010]. We now present the two main results for the game $(N, C_{fix})$ analysis in Theorems 2.1 and 2.2. We start with the most profitable coalition structure.

**Theorem 2.1.** The service pooling game $(N, C_{fix})$ is a strictly subadditive game, and the grand coalition is the most profitable coalition structure.

*Proof.* For the total cost function $C_{fix}(U) = C_{h,U} + C_{w,U}$, the first term $C_{h,U} = c_h \mu_U$ is additive, and the second one $C_{w,U} = c_w L_{q,U}$ is strictly subadditive, based on Proposition 2.1. It is then clear that $C_{fix}$ is strictly subadditive, so $(N, C_{fix})$ is a strictly subadditive game, and any splitting of the grand coalition implies an additional congestion cost for the entire set $N$. Thus, the grand coalition $N$ is the most profitable coalition structure for the game $(N, C_{fix})$. □

Subadditivity is a necessary condition required for the formation of the grand coalition. It states that $N$ is the most profitable coalition. Theorem 2.1 states that there is always a benefit if service providers share their service capacities as a 'super-server' with a fixed service capacity. In the context of resource pooling, each participant could save money from the coalition congestion cost. Although the profitability is justified by Theorem 2.1, the existence of stable cost allocations has not been confirmed yet. In order to motivate every provider to join the grand coalition, our interest now is to find a stable allocation rule to share the reduced congestion cost. One of the important properties for the stable cost allocation analysis is the concavity as given in the following definition.

**Definition 2.1.** A TU-game $(N, v)$ is concave if for any pair of subsets $\emptyset \subseteq U \subset T \subset N$, and any player $l \in N \backslash T$, $v(U \cup \{l\}) - v(U) \geq v(T \cup \{l\}) - v(T)$.

The Shapley value, introduced by Shapley in 1952 [Shapley, 1952], is an important allocation concept in cooperative game theory. It provides us with a well known fair

allocation rule for TU-games, but it could not always stay in the core. For a TU-game $(N, v)$, the Shapley value is given by

$$sh_i = \sum_{U \subseteq N \setminus \{i\}} \frac{|U|!(|N| - |U| - 1)!}{|N|!} [v(U \cup \{i\}) - v(U)], i \in N, \tag{2.5}$$

where $v(U \cup \{i\}) - v(U)$ is the marginal cost of provider $i$ as the last player joining the coalition $U$. If $(N, C_{fix})$ is concave, the non-emptiness of the core and the stability of the Shapley value would be ensured. Unfortunately, the following example illustrates the non-concavity of the game $(N, C_{fix})$ and the instability of the Shapley value as well as the classical proportional allocation rules.

Consider the case where $c_h = 0$ (because of the additivity of capacity cost, this assumption does not affect the results of cooperative games), $c_w = 1$ and $f = 0.2$, a set $N = \{1, 2, 3\}$ of three service providers, and the remaining parameters as defined in Table 2.1.

| Players | Parameters | | |
|---|---|---|---|
| | $\lambda_i$ | $\mu_i$ | $C_i$ |
| 1 | 9 | 10 | 4.86 |
| 2 | 5 | 10 | 0.3 |
| 3 | 2 | 10 | 0.03 |

**Table 2.1:** *3 players pooling game*

We use the same arrival and service rates of the example in [Yu et al., 2015]. There are big differences among service utilizations of each player under this situation. By choosing $U = \{1\}$, $T = \{1, 3\}$ with $l = 2$, we obtain $C_{fix}(U \cup \{l\}) - C_{fix}(U) = -3.88 < C_{fix}(T \cup \{l\}) - C_{fix}(T) = -0.03$, meaning that the game $(N, C_{fix})$ is not concave in this setting.

For this example, we have $sh_1^{fix} = 1.88$, $sh_2^{fix} = -0.54$, $sh_3^{fix} = -0.97$. We denote the demand and the contribution by $\lambda_i$ and $\mu_i - \lambda_i$, respectively. The relative contribution values are calculated as $(\mu_i - \lambda_i)/\lambda_i$, which give $(0.11, 1, 4)$ for $N$. The ordering of these values is consistent with that of the profit rates, i.e., $[C(\{i\}) - sh_i^{fix}]/C(\{i\})$ calculated

by $sh^{fix}$. In Table 2.2, the total cost of every coalition and the corresponding cost of the Shapley value allocation are both computed.

| $U$ | $C_{fix}(U)$ | $\sum_{i\in U} sh_i^{fix}$ |
|---|---|---|
| {1,2,3} | 0.37 | 0.37 |
| {1,2} | 0.98 | 1.34 |
| {2,3} | 0.11 | $-1.52$ |
| {1,3} | 0.40 | 0.91 |

**Table 2.2:** *Coalition cost and distributed cost by sh*

We have $C_{fix}(\{1,2\}) = 0.98 < sh_1^{fix} + sh_2^{fix} = 1.88 + (-0.54) = 1.34$, also, $C_{fix}(\{1,3\}) = 0.4 < sh_1^{fix} + sh_3^{fix} = 1.88 + (-0.97) = 0.91$. The Shapley value allocation is therefore not stable in this case, although the allocation captures the contribution of each player. The general proportional allocation rule $\varphi_i^p = p_i C_{fix}(N) / \sum_{j\in N} p_j$, depending on the initial individual service capacity ($p_i = \mu_i$) or the own customer arrival rate ($p_i = \lambda_i$), also does not guarantee the stability of $N$.

In order to keep all players staying in the coalition, we should find at least one stable cost allocation if it exists. Unfortunately, it is very hard to give an explicit cost allocation rule for the game $(N, C_{fix})$. However, we prove the existence of the stable cost allocation rules. We employ the "Bondareva-Shapley Theorem" [Bondareva, 1963] (B-S Theorem), which is known for the non-empty core proofs of TU-games: "A TU-game $(N, v)$ has a non-empty core if and only if it is balanced". The following definitions are relevant notions of balancedness.

**Definition 2.2.** A collection $B$ on $N$ ($B$ consists of sub-coalitions of $N$) is a *balanced collection*, if there exist *weights* $\beta_U \in [0,1]$ such that $\sum_{U\in B} \beta_U \mathbf{1}_U = \mathbf{1}_N$. This equation is equivalent to $\sum_{U\ni i} \beta_U = 1$, for any $i \in N$.

**Definition 2.3.** A cost TU-game $(N, v)$ is a *balanced game*, if for any balanced collection $B$ on $N$, we have $v(N) \le \sum_{U\in B} \beta_U v(U)$ ($v(N) \ge \sum_{U\in B} \beta_U v(U)$ for profit games).

For the balancedness proof of the game $(N, C_{fix})$, we use the following proposition.

**Proposition 2.2.** The expected waiting time $W_q(\lambda_U, \mu_U)$ in the queue is a decreasing and convex function in $\mu_U$ with fixed $\lambda_U$, for $\mu_U > \lambda_U$.

*Proof.* We have $\partial W_q/\partial \mu_U = -\lambda_U(1+f)(2\mu_U - \lambda_U)/[2\mu_U^2(\mu_U - \lambda_U)^2] < 0$ and $\partial^2 W_q/\partial \mu_U^2 = \lambda_U(1+f)[\lambda_U^2 + 3\mu_U(\mu_U - \lambda_U)]/[\lambda_U^3(\lambda_U - \lambda_U)^3] > 0$, for $\mu_U > \lambda_U$. So, $W_q(\mu_U)$ is decreasing and convex in $\mu_U$. □

With Proposition 2.2, we are now prepared to prove the non-emptiness of the core for the game $(N, C_{fix})$ as given in Theorem 2.2.

**Theorem 2.2.** The service pooling game $(N, C_{fix})$ has a non-empty core, and there are infinitely many solutions in the core if $n > 1$.

*Proof.* The game $(N, C_{fix})$ could be divided into two games: the game $(N, C_h)$ with $C_h(U) = c_h\mu_U$, which is a linear game resulting in a single core allocation for any $i \in N$, $\varphi_i = c_h\mu_i$, and the game $(N, C_w)$ with $C_w(U) = c_w L_q(\lambda_U, \mu_U) = c_w\lambda_U W_q(\lambda_U, \mu_U)$. It is clear that the only core allocation of $(N, C_{fix})$ is the sum of the two games'. Now, we will prove that the game $(N, C_w)$ has a non-empty core using the B-S Theorem.

For any balanced collection $B$ on $N$, we have

$$\begin{aligned}
C_w(N) &= d\lambda_N W_q(\lambda_N, \mu_N) \\
&= c_w\lambda_N W_q(\lambda_N, \sum_{U\in B} \beta_U\mu_U \frac{\lambda_N}{\lambda_U}\cdot\frac{\lambda_U}{\lambda_N}) && (2.6)\\
&\le c_w\lambda_N \cdot \sum_{U\in B} \beta_U \frac{\lambda_U}{\lambda_N} W_q(\lambda_N, \mu_U\frac{\lambda_N}{\lambda_U}) && (2.7)\\
&= \sum_{U\in B} \beta_U c_w\lambda_U W_q(\lambda_N, \mu_U\frac{\lambda_N}{\lambda_U}) \\
&= \sum_{U\in B} \beta_U c_w\lambda_U [\frac{1+f}{2}\lambda_N^{-1}\frac{1}{\rho_U^{-1}(\rho_U^{-1}-1)}] \\
&< \sum_{U\in B} \beta_U c_w\lambda_U W_q(\lambda_U, \mu_U) = \sum_{U\in B}\beta_U C_w(U). && (2.8)
\end{aligned}$$

From the definition of a balanced collection, there is $\sum_{U\in B}\beta_U\mu_U = \mu_N$ to guarantee the equality in (2.6). Meanwhile, the inequality in (2.7) holds by the convex property of

$W_q(\mu_U)$ in Proposition 2.2 and $\sum_{U\in B}\beta_U\lambda_U = \lambda_N$ from Definition 2.2. Since $U$ is a subset of $N$, with $\lambda_N^{-1} < \lambda_U^{-1}$, the inequality in (2.8) holds.

Thus, the game $(N, C_w)$ is a balanced game. According to the B-S Theorem, the game $(N, C_w)$ has a non-empty core. Using Lemma A.2 of [Karsten et al., 2015b], we could simply state that: if $n > 1$, the game $(N, C_{fix})$ has infinitely many core allocations. The proof of the theorem is completed. □

From Theorem 2.2, we could conclude that the game $(N, C_{fix})$ has always a core allocation to maintain the stability of the grand coalition. For further results on cost-sharing, an explicit numerical solution of the game $(N, C_{fix})$ could be computed through mathematical programming, using for example, the Equal Profit Method in [Frisk et al., 2010] or Nucleolus computing [Schmeidler, 1969]. However, this does not allow for an explicit characterization of the allocations. In the next part, we treat this game with several programming methods.

### 2.4.2. Cost allocation rules

When analyzing a TU-game, the challenging problem is to provide a mechanism to motivate all players to join the profitable coalition. An *allocation* $\boldsymbol{\varphi} = \{\varphi_i, i \in N\} \in \mathbb{R}^n$ for $(N, C)$ provides a way for sharing the cost over the players. If $\sum_{i\in N}\varphi_i = C(N)$, $\boldsymbol{\varphi}$ is *efficient*. If $\varphi_i \leq C(i)$ for any $i \in N$, $\boldsymbol{\varphi}$ is *individual rational*. For any $U \subseteq N$, when $\sum_{i\in U}\varphi_{N,i} \leq C(U)$, $\boldsymbol{\varphi}$ corresponds to the *coalitional rationality*. If $\boldsymbol{\varphi}$ is justified for all three properties above, this allocation is *stable* for the game $(N, C)$. Moreover, all the stable allocations form the *core* of a TU-game.

#### Shapley-value

The *Shapley-value*, introduced by Shapley in 1952 [Shapley, 1952], is a popular allocation concept in cooperative game theory. It provides us with a well-known fair allocation rule for cooperative games, but there is no general property to keep the stability of the grand

coalition. The Shapley-value is defined as the average marginal cost of each cooperative subset for each participant, and it is given in Equation (2.5).

Shapley-value is the unique allocation rule satisfying the four desirable properties: anonymity, efficiency, additivity and dummy player property. In order to verify the Shapley-value is a core allocation for a cost cooperative game $(N, v)$, it is sufficient to test the concavity of the characteristic function $v$, which has been proved by [Shapley, 1971]. Unfortunately, the game $(N, C_{fix})$ defined here is non-concave and the Shapley-value couldn't guarantee the stability of grand coalition in general.

**Tau-value**

The set of cost allocations, defined by the lower bound $M_i(N, C_{fix}) = C_N - C_{N\setminus\{i\}}$ and the upper bound $m_i(C_{fix}) = max_{U:i\in U}\{C_U - \sum_{j\in U\setminus\{i\}} M_j(N, C_{fix})\}$ of the shared cost, includes all the core allocations. The $\tau$-value, defined by Tijs in 1981 [Tijs, 1981], is a cost allocation defined as

$$\varphi_i^\tau = \alpha m_i(C_{fix}) + (1-\alpha)M_i(N, C_{fix}), \tag{2.9}$$

with the unique $\alpha \in [0, 1]$ calculated by the efficiency of $\varphi^\tau$. It is a special linear combination of $M_i(N, C_{fix})$ and $m_i(C_{fix})$. For two player games, the $\tau$-value is equal to the Shapley-value and it presents a stable cost allocation for all quasi-balanced two player games. Although we could not give an explicit demonstration of its stability, the $\tau$-value proposes stable results in the following experiments.

**Nucleolus**

The *excess* of a cost allocation $\varphi$ for a coalition $U \subseteq N$ is defined as

$$e_{\varphi,U} = C_U - \sum_{i\in U} \varphi_i, \tag{2.10}$$

which is been used to measure *unhappiness* of the coalitions by the *lexicographic ordering*

comparison for all possible allocations. When the coalition cost is allocated by $\varphi$, $U$ is more satisfied with a higher $e_{\varphi,U}$. Based on the concept of *minimized maximum unhappiness*, the *Nucleolus* was introduced by Schmeidler in [Schmeidler, 1969]. Although Schmeidler didn't define it explicitly, its uniqueness has been proved by Driessen in 1969 [Driessen, 1988]. It is a stable allocation rule for all TU-games with a non-empty core, and coincides with dummy player property, zero independent and reduced-game property.

**Equal Profit Method**

Another interesting definition of "fair" is the *equal profit* concept, which defined by Frisk et al. in 2010 [Frisk et al., 2010]. They describe the cost TU-games as a linear programming problem to minimize the gap of the relative savings $r_i = (C_i - \varphi_i)/C_i$.

$$\begin{aligned} &\min f(\varphi) \\ &\quad \text{s.t } f(\varphi) \geq \max\{r_i - r_j\}, \forall i, j \in N \\ &\qquad \textstyle\sum_{i \in U} \varphi_i \leq C_U, \forall U \subset N \\ &\qquad \textstyle\sum_{i \in N} \varphi_i = C_N, \end{aligned} \tag{2.11}$$

where the results $\varphi^{EPM}$ defined as the core allocation calculated by *Equal Profit Method* (EPM). It seems to be a "fair" and stable allocation rule considering the most closed relative savings for each individual, despite the different contribution of each participant to the grand coalition $N$. This "fair" definition just considers the individual factor. In some cases, the participant, which has a low individual payment, might do not want to join the coalition in order to protect its individual information, although it could get a same level relative saving with others.

**EPM based on Contribution Weights**

Now, we define the *relative contribution weights* $w_i = \lambda_i/(\mu_i - \lambda_i)$ for each participant in our service pooling game. The customer arrival rate $\lambda_i$ describes the individual requirement of each participant, and the system idle capacity $\mu_i - \lambda_i$ presents the individual contribution to the collaboration. Then, we propose the new relative saving formula as

$$rw_i = w_i * \frac{C_i - x_i}{C_i} = \frac{\lambda(C_i - x_i)}{(\mu - \lambda)C_i}. \tag{2.12}$$

Thus, we get the *EPM with Contribution Weights* (EPMCW) as

$$\begin{aligned} &\min f(\varphi) \\ &\text{s.t } f(\varphi) \geq \max\{rw_i - rw_j\}, \forall i, j \in N \\ &\quad \textstyle\sum_{i \in U} \varphi_i \leq C_U, \forall U \subset N \\ &\quad \textstyle\sum_{i \in N} \varphi_i = C_N. \end{aligned} \tag{2.13}$$

With the constraints defined in equation (2.13), EPMCW presents a stable allocation rule by solving the linear programming problem above, if the core is not empty. The contribution of each provider has been considered in the *relative contribution weights* $w_i$, and the results are easily controllable by $w_i$ with different definitions, e.g., $w_i^* = \lambda_i^2/(\mu_i - \lambda_i)$.

### 2.4.3. Numerical results and analysis

We illustrate the previous concepts in three typical service pooling cases of 6 service companies, which have equal individual service capacities. To do so, we consider the three following sets of data with $c_h \in \{0, 1\}$ and $f \in \{0, 1, 4\}$ (the variation of the two system parameters $\{c_h, c_w\}$ have similar impacts on the results). The first case presents a set of companies with low server utilizations. It means that all the companies in this set are not very efficient. There are both the less efficient companies and the busy companies

in the second case. And the third one only consists of the busy companies. The initial parameters are listed in Table 2.3.

| Initial data | | | | | Customer arrival rates | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\mu$ | $c_h$ | $c_w$ | $f$ | **No.** | 1 | 2 | 3 | 4 | 5 | 6 |
| 10 | $\{0,1\}$ | 2 | $\{0,1,4\}$ | **1** | 2 | 3 | 3 | 4 | 3.5 | 4 |
| - | | | | **2** | 7 | 9 | 8 | 7 | 8.5 | 9.5 |
| - | | | | **3** | 2 | 7 | 4 | 7.5 | 9 | 3 |

**Table 2.3:** *Customer arrival rates and system parameters*

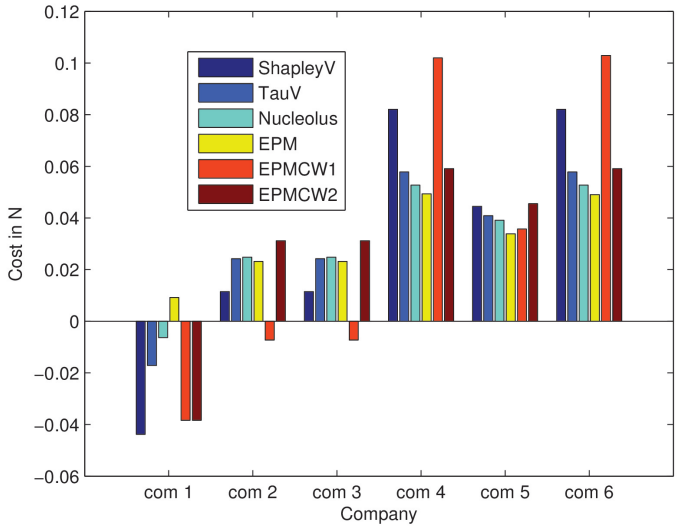



**Figure 2.2:** *Cost allocations of case 1 under $c_h = 0$ by Shapley-value, Tau-value, Nucleolus, EPM, EPMCW1 and EPMCW2*

Figures 2.2, 2.3 and 2.4 reveal the distributed costs, using different preceding concepts mentioned in 2.4.2 of the three cases above with $c_h = 0$, i.e., the quality driven case. We use two different relative contribution rates for EPMCW calculation: $w_i$ for EPMCW1 and $w_i^*$ for EPMCW2. From these figures, we find that the results given by the $\tau$-value and the nucleolus are very close, especially in the third case. The Shapley-value reflects the contributions of each provider for all the possible coalitions. Unfortunately, the corresponding allocations are not stable in all the three cases, e.g., $sh_1 + sh_5 = 1.2811 \geq C_{\{1,5\}} = 0.8067$ in the case 3.

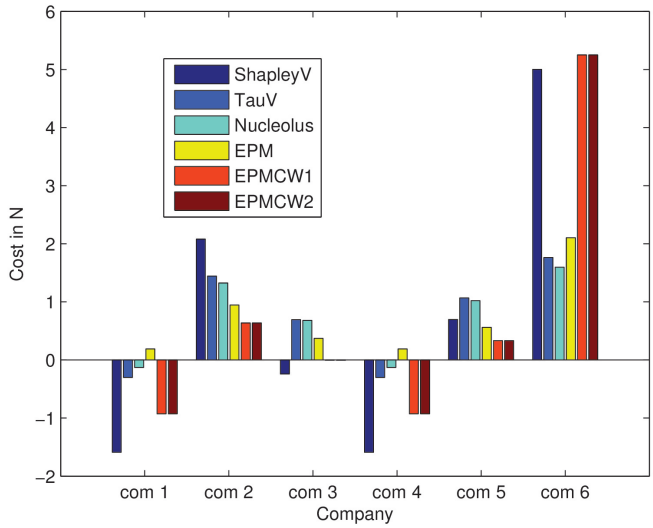



**Figure 2.3:** *Cost allocations of case 2 under $c_h = 0$ by Shapley-value, Tau-value, Nucleolus, EPM, EPMCW1 and EPMCW2*

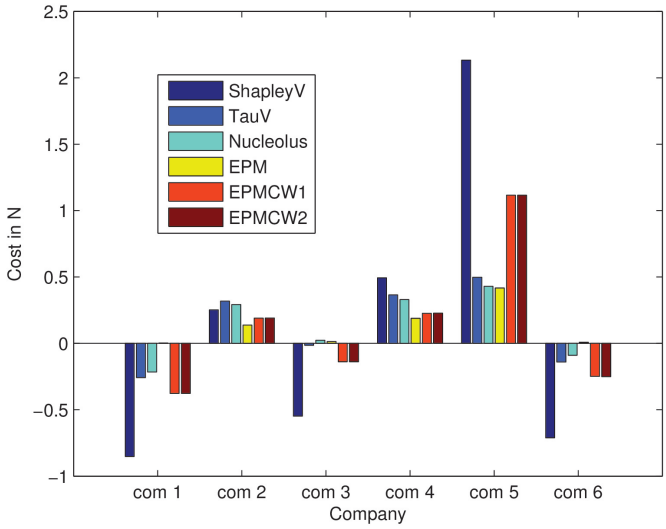



**Figure 2.4:** *Cost allocations of case 3 under $c_h = 0$ by Shapley-value, Tau-value, Nucleolus, EPM, EPMCW1 and EPMCW2*

EPM provides a stable allocation rule, which is a little far from the others for several companies, particularly for the company with relative large or small contribution to the coalitions. The goal or the "fair" defined by EPM is to minimize the gap between the relative earnings of players, and constraints of this programming consist of all stable re-

quirements. Thus, the companies with a special contribution may not satisfy the similar relative saving. For example, the first company in Figure 2.7, which has a low individual operating cost, will pay nothing using EPM allocation, but it could earn money in other allocations.

| $c_h = 0$ | Shapley | Tau-v | Nucl. | EPM | E-CW1 | E-CW2 |
|---|---|---|---|---|---|---|
| Com 1 | 9.91% | 7.36% | 6.33% | 4.85% | 9.40% | 9.40% |
| Com 2 | 13.64% | 12.43% | 12.37% | 12.52% | 15.43% | 11.76% |
| Com 3 | 13.64% | 12.43% | 12.37% | 12.52% | 15.43% | 11.76% |
| Com 4 | 22.72% | 25.04% | 25.53% | 25.85% | 20.82% | 24.92% |
| Com 5 | 17.35% | 17.70% | 17.87% | 18.38% | 18.19% | 17.26% |
| Com 6 | 22.72% | 25.04% | 25.53% | 25.87% | 20.74% | 24.92% |
| $d_\varphi$ | 4.27% | 5.93% | 6.31% | 6.70% | 3.25% | 5.70% |
| Stability | N | Y | Y | Y | Y | Y |
| $c_h = 0.5$ | Shapley | Tau-v | Nucl. | EPM | E-CW1 | E-CW2 |
| Com 1 | 9.91% | 7.37% | 6.33% | 9.40% | 9.40% | 9.40% |
| Com 2 | 13.64% | 12.43% | 12.37% | 11.76% | 11.76% | 11.76% |
| Com 3 | 13.64% | 12.43% | 12.37% | 11.76% | 11.76% | 11.76% |
| Com 4 | 22.73% | 25.04% | 25.53% | 24.92% | 24.92% | 24.92% |
| Com 5 | 17.35% | 17.70% | 17.86% | 17.25% | 17.25% | 17.25% |
| Com 6 | 22.73% | 25.04% | 25.53% | 24.92% | 24.92% | 24.92% |
| $d_\varphi$ | 4.27% | 5.93% | 6.31% | 5.70% | 5.70% | 5.70% |
| Stability | N | Y | Y | Y | Y | Y |

**Table 2.4:** *Profit-sharing $sp_i$ for case 1 with $c_h = \{0, 1\}$ and $f = 4$*

EPMCW1 and EPMCW2 propose similar allocations in Figures 2.3 and 2.4, but they propose different results in Figure 2.2. In these figures, we find that the results of EPMCW1 and EPMCW2 are closer to the Shapley value than the others. For more detailed comparison, we consider the profit distribution, which is defined as $sp_{\varphi,i} = (C_i - \varphi_i)/(\sum_{i\in N} C_i - C_N)\%$. Considering the fairness issue, we denote the saving deviation by $d_\varphi = \sum_{i\in N} |sp_{\varphi,i} - 1/n|/n$. Therefore, the less $d_\varphi$ presents the fairer in terms of equal savings. Some data of the case 1 is listed in Table 2.4.

In Table 2.4, it is obvious that EPM and EPMCW are not additive allocation methods.

When $c_h = 0$, the allocation using EPMCW1 is the most fair allocation in terms of equal savings. While the relative saving declines, $c_h = 0.5$, the difference between EPM and EPMCWs reduces. If the relative saving is small enough, e.g., $c_h = 1$ with $r_N = 6.69\%$, the two allocations given by EPMCW and the allocation of EPM are very similar.

## 2.5. Service pooling game with optimized service capacity

We consider here the case where the service capacity could be optimized in order to minimize the total operating cost. This section is separated into four parts. First, it discusses the properties of the optimal service rate in M/GI/1 service systems. Second, it characterizes and investigates the service pooling game with the optimized service capacity. The last two parts analyze and compare between two special stable allocation rules for this game.

### 2.5.1. Optimal service rate in M/GI/1 systems

For certain situations, the service capacity sometimes could be adjusted as required. We assume here that the service capacity of each individual provider or service coalition is a continuous variable. The optimal service rate for an M/GI/1 service system with a given customer arrival rate $\lambda$ is defined as

$$\mu^* = argmin\{c_h\mu + c_w\frac{\lambda^2(1+f)}{2\mu(\mu-\lambda)}|\mu > \lambda\}. \tag{2.14}$$

It is however hard to obtain a closed-form expression of $\mu^*$. We rewrite the minimization problem above using the server utilization $\rho^* = \lambda/\mu^*$ as

$$\rho^* = argmin\{c_h\lambda\rho^{-1} + \frac{c_w(1+f)}{2}\frac{1}{\rho^{-1}(\rho^{-1}-1)}|\rho \in (0,1)\}. \tag{2.15}$$

For a given $\lambda$, it is clear that the total cost $C$ is a single-variable function in $\rho$. We state some useful properties in Lemma 2.1 related to the optimal service rate $\mu^*$ and the optimal server utilization $\rho^*$.

**Lemma 2.1.** The following holds.

*(i)* There exists a unique optimal service rate $\mu^*$, for a given $\lambda$;

*(ii)* The optimal server utilization $\rho^*$ is increasing in $\lambda$;

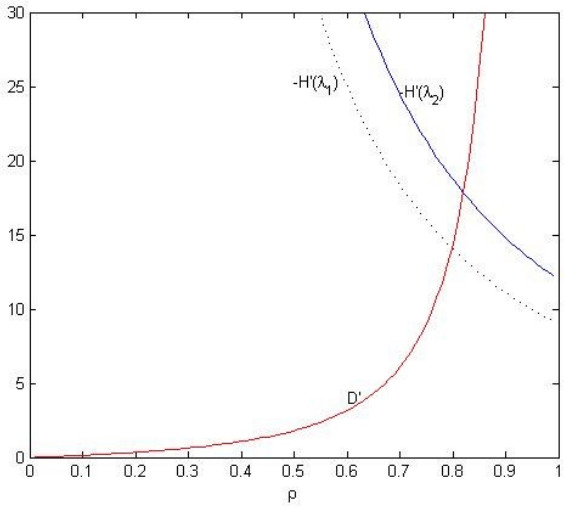
*(iii)* The optimal server utilization $\rho^*$ is decreasing in $f$.

*Proof.* The cost function $C(\cdot)$ defined by $\{c_h, c_w, f, \lambda\}$ in Equation (2.4) consists of two parts: the service capacity cost $C_h = c_h\lambda\rho^{-1}$, with $\partial C_h/\partial\rho = -c_h\lambda\rho^{-2} \in (-\infty, -c_h\lambda)$, is a decreasing function in $0 < \rho < 1$; and the system congestion cost $C_w = [c_w(1+f)]/[2\rho^{-1}(\rho^{-1}-1)]$, with $\partial C_w/\partial\rho = [c_w(1+f)(2\rho^{-1}-1)]/[2(\rho^{-1}-1)^2] \in (0, +\infty)$, is an increasing function of $\rho \in (0,1)$. Meanwhile, $-(\partial^2 C_h)/\partial\rho^2 = -c_h\lambda\rho^{-3} < 0$ and $\partial^2 C_w/\partial\rho^2 = [c_w(1+f)\rho^{-3}]/(\rho^{-1}-1)^2 > 0$. Then, the two first-order differential equations, $-\partial C_h/\partial\rho$ and $\partial C_w/\partial\rho$, intersect only once on $(0,1)$. Thus, $\rho^*$ is the intersection $-\partial C_h/\partial\rho = \partial C_w/\partial\rho$, which is unique in the definition interval of $\rho$. It means that the optimal service rate $\mu^*$ exists and is unique in its definition interval $(\lambda, +\infty)$. This proves the first part of Lemma 2.1.

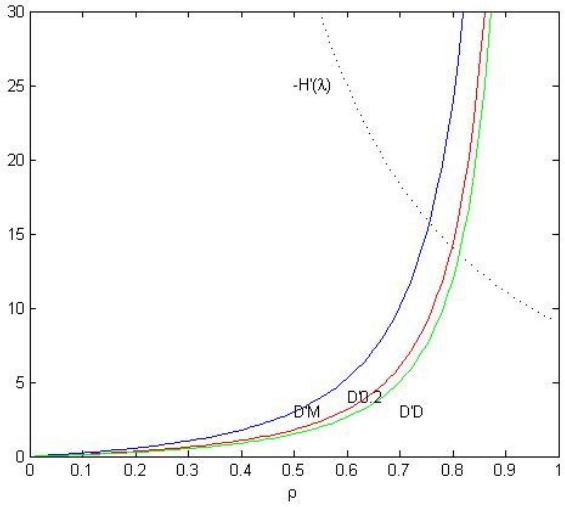
If the customer arrival rate $\lambda$ varies, the partial differential equation $\partial C_w/\partial\rho$ is defined by $\{c_h, c_w, f\}$. Because of $-(\partial^2 C_h)/\partial\rho^2 = -c_h\lambda\rho^{-3} < 0$, the slope of $-\partial C_h/\partial\rho$ decreases in $\lambda$. This is to say that the larger is $\lambda$, the slower is the decrease of $-\partial C_h/\partial\rho$. Therefore, the intersection is obtained by a larger server utilization $\rho^*$ and (ii) of Lemma 2.1 is proved. The last part of Lemma 2.1 can be proved using a similar analysis, by varying the variability of service time $f$ for a fixed arrival rate $\lambda$. This finishes the proof of the lemma. □

The properties stated in Lemma 2.1 are illustrated in Figure 2.5. An intuitive explanation for (ii) of Lemma 2.1 is as follow. The necessary service capacity for one unit demand, which minimizes the system total cost, decreases with the increase of the service size. Statement (iii) of Lemma 2.1 explains that an extra service capacity is required to deal with the impact of service variability. If $f = 0$, the queueing model becomes an M/D/1 queue, for which the minimum total cost is achieved with the fewest necessary service capacity for one unit demand.

Given the monotonicity of $\rho^*$ in $\lambda$, it is interesting to study the monotonicity of $\mu^*$




(a) $-\partial C_h/\partial\rho$ and $\partial C_w/\partial\rho$ with $\lambda_1 < \lambda_2$




(b) $-\partial C_h/\partial\rho$ and $\partial C_w/\partial\rho$ with $f = \{1, 0.2, 0\}$

**Figure 2.5:** *Slope of $C_h$ and $C_w$ with different $\lambda$ and $f$*

in $\lambda$. This is provided in (i) of Lemma 2.2. As we mentioned in the previous section, the game $(N, C_{fix})$ is not monotone, i.e., $D_s$ could be reduced or increased by a new joining player. Let us denote by $C^*$ the total operating cost with the optimized service capacity $\mu^*$, which is a single-variable function of the customer arrival rate $\lambda$, and $\mu^*$ is

the argument resulting from Equation (2.14). We have

$$C^*(\lambda) = min\{C|\lambda \in \mathbf{R}^+\} = C(\lambda, \mu^*(\lambda)). \tag{2.16}$$

We show a monotonicity result of $C^*$ in (ii) of Lemma 2.2.

**Lemma 2.2.** The following holds.

*(i)* The optimal service rate $\mu^*$ is increasing in $\lambda$;

*(ii)* The optimal cost $C^*$ is increasing in $\lambda$.

*Proof.* We have $\partial L_q/\partial\mu < 0$ and $\partial(\partial L_q/\partial\mu)/\partial\lambda < 0$ with $\lambda < \mu$. It means that the queue length is faster reduced in $\mu$ with a high $\lambda$, which corresponds to the advantage of service pooling. For two systems with $\lambda_1 > \lambda_2$, we have

$$\frac{\partial L_q(\lambda_1)}{\partial\mu} < \frac{\partial L_q(\lambda_2)}{\partial\mu}, \text{and } \frac{\partial L_q(\lambda_1, \mu_1^*)}{\partial\mu} = \frac{\partial L_q(\lambda_2, \mu_2^*)}{\partial\mu} = \frac{-2c_h}{c_w(1+f)}.$$

So, $\mu_1^* > \mu_2^*$. The proof of the first part is finished. For the second part, we have

$$\begin{aligned}\frac{\partial C^*}{\partial\lambda} &= c_h\frac{\partial\mu^*}{\partial\lambda} + \frac{c_w(1+f)}{2} * \frac{\lambda(\mu^* - \lambda * \partial\mu^*/\partial\lambda)(2\mu^* - \lambda)}{\mu^{*2}(\mu^* - \lambda)^2} \\ &= c_h\frac{\partial\mu^*}{\partial\lambda} + \frac{c_w(1+f)}{2} * \frac{\lambda(2\mu^* - \lambda)}{(\mu^* - \lambda)^2} * \frac{\partial\rho^*}{\partial\lambda}.\end{aligned}$$

Using statement (ii) of Lemma 2.1 and the result of the first part, we have $\partial C^*/\partial\lambda > 0$. This completes the proof of the lemma. □

From (i) of Lemma 2.2, we could conclude that $\mu^*$ always increases with a new joining player even though the service capacity cost is much more higher than the congestion cost per unit time, i.e., $c_h >> c_w$. Furthermore, statement (ii) of Lemma 2.2 shows that $C^*$ always increases due to the addition of new providers.

### 2.5.2. Service pooling game with optimized service capacity

Using Lemma 2.1, we can construct a service pooling TU-game with the optimized service capacity for M/GI/1 service systems. Let $C_{opt}(\cdot) : 2^N \to \mathbf{R}$ be the total cost for each

coalition $S \subseteq N$ using $\mu_U^*$ defined with $\{c_h, c_w, f\}$ and $\lambda_U = \sum_{i \in U} \lambda_i$. We have

$$C_{opt}(\emptyset) = 0, C_{opt}(U) = C^*(\lambda_U), \text{for any } \emptyset \subset U \subseteq N. \tag{2.17}$$

Similarly to the game $(N, C_{fix})$ in the previous section, the game $(N, C_{opt})$ defined with the optimized service capacity is also a TU-game. It is clear that this system has less expenses than the previous situation. Furthermore, each service provider has only one single individual parameter, i.e., the customer arrival rate $\lambda_i$. This makes the search for a stable cost allocation rule easier than the case with a fixed $\mu$.

**Theorem 2.3.** The service pooling game $(N, C_{opt})$ is a strictly subadditive game, and the grand coalition is the most profitable coalition structure.

*Proof.* With the subadditive property of the game $(N, C_{fix})$, for any pair of subset $\emptyset \subset U, T \subset N$ with $U \cup T = \emptyset$, we may write

$$\begin{aligned} C_{opt}(U \cup T) &= C^*(\lambda_{U \cup T}) = min\ C_{U \cup T}(\lambda_{U \cup T}) \\ &\leq C_{fix}(\lambda_{U \cup T}, \mu_U^* + \mu_T^*) \\ &< C_{fix}(\lambda_U, \mu_U^*) + C_{fix}(\lambda_T, \mu_T^*) \\ &= C^*(\lambda_U) + C^*(\lambda_T) = C_{opt}(U) + C_{opt}(T). \end{aligned}$$

Thus, the game $(N, C_{opt})$ is strictly subadditive and no other coalition structure is more profitable than the grand coalition $N$. □

In contrast to the game $(N, C_{fix})$, we prove that the game $(N, C_{opt})$ is concave with any set of $\{\lambda_1, \cdots, \lambda_n\}$ in the following theorem.

**Theorem 2.4.** The service pooling game $(N, C_{opt})$ is a concave game.

*Proof.* To prove the concavity of $C_{opt}(\cdot)$, we use the inequality in Definition 2.1. Consider a pair of coalitions $\emptyset \subseteq U \subset T \subset N$ and any $l \in N \setminus T$. It is then sufficient to show that

$$C_{opt}(U \cup \{l\}) - C_{opt}(U) \geq C_{opt}(T \cup \{l\}) - C_{opt}(T). \tag{2.18}$$

Using Equations (2.16) and (2.17), Equation (2.18) is equivalent to

$$C^*(\lambda_{U\cup\{l\}}) - C^*(\lambda_U) \geq C^*(\lambda_{T\cup\{l\}}) - C^*(\lambda_T). \tag{2.19}$$

Since $U \subset T$ and $l \in N \setminus T$, $\lambda_U < \lambda_T$ and $\lambda_{U\cup\{l\}} = \lambda_U + \lambda_l \leq \lambda_{T\cup\{l\}} = \lambda_T + \lambda_l$. We assume that $\triangle_\lambda$ is an infinitesimal of the customer arrival rate $\lambda$. Consider that $m\triangle_\lambda = \lambda_l$, with $m \in \mathbf{N}^+$. Therefore, the right-hand side of Equation (2.19) can be considered as the integral of the partial differential of the total cost function $C^*$ from $\lambda_U$ to $\lambda_{U\cup\{l\}}$. This also holds for the left-hand side. Because the optimal service rate $\mu^*$ varies with $\lambda$, we compare the differential in every infinitesimal interval of $\lambda_l$.

$$\begin{aligned}
&C^*(\lambda_{U\cup\{l\}}) - C^*(\lambda_U) \\
=&\sum_{i\in[i:m]} [C^*(\lambda_U + \lambda_l - (i-1)\triangle_\lambda) - C^*(\lambda_U + \lambda_l - i\triangle_\lambda)] \\
=&\sum_{i\in[i:m]} [\triangle_\lambda \frac{\partial C(\lambda_U + \lambda_l - i\triangle_\lambda | \rho^*_{\lambda_U+\lambda_l-i\triangle_\lambda})}{\partial\lambda}] \\
=&\sum_{i\in[i:m]} h\rho^{*-1}_{\lambda_U+\lambda_l-i\triangle_\lambda}\triangle_\lambda \\
\geq&\sum_{i\in[i:m]} h\rho^{*-1}_{\lambda_T+\lambda_l-i\triangle_\lambda}\triangle_\lambda \\
=&C^*(\lambda_{T\cup\{l\}}) - C^*(\lambda_T).
\end{aligned} \tag{2.20}$$

Because $\lambda_U + \lambda_l - i\triangle_\lambda \leq \lambda_T + \lambda_l - i\triangle_\lambda$ for any $i$ in $[1:m]$ , the inequality in (2.20) holds from the inequality $\rho^*_{\lambda_U+\lambda_l-i\triangle_\lambda} \leq \rho^*_{\lambda_T+\lambda_l-i\triangle_\lambda}$, which is based on the statement (ii) of Lemma 2.1 about the optimal server utilization $\rho^*$. Thus, we obtain Equation (2.19) by $\rho^{*-1}_{\lambda_U+\lambda_l-i\triangle_\lambda} \geq \rho^{*-1}_{\lambda_T+\lambda_l-i\triangle_\lambda}$, and the game $(N, C_{opt})$ is concave. This finishes the proof of the theorem. □

As an immediate consequence of Theorem 2.4, the game $(N, C_{opt})$ is totally balanced. There are infinitely many solutions in the core if $n > 1$. Being a concave cost game, the gains of joining a coalition for the same service provider decrease as the coalition grows, and all the core allocations could be written as a convex combination of marginal

contribution vectors (the corresponding convex value game has been perfectly explained by Shapley at 1971 in [Shapley, 1971]).

### 2.5.3. Cost allocation rules for the service pooling game $(N, C_{opt})$

From the individual point of view, it is very important to design a pooling mechanism, which motivates all service providers to join the grand coalition. A fair and efficient cost allocation rule is necessary among players to avoid splitting incentives of sub-coalitions. In what follows, we focus on the Shapley value and the proportional allocation rules.

#### Shapley-value

As mentioned for the game $(N, C_{fix})$, it is sufficient to verify the concavity of the characteristic function $v$ [Shapley, 1971]. Since the game $(N, C_{opt})$ is proved as a concave cost game in Theorem 2.4, the next result follows.

**Proposition 2.3.** The Shapley value *sh* provides a stable cost allocation rule for the service pooling game $(N, C_{opt})$.

$$sh_i^{opt} = \sum_{U \subseteq N \setminus \{i\}} \frac{|U|!(|N| - |U| - 1)!}{|N|!} [C_{opt}(U \cup \{i\}) - C_{opt}(U)]. \tag{2.21}$$

#### Proportional allocation based on customer arrival rates

The higher is the number of service providers joining the pooling system, the more complex is the computation of the Shapley value (e.g., if $n = 20$, than there are $2^n - 1$ possible coalitions in total, i.e., millions of optimal service rates should be worked out in the preliminary computing). We therefore need to have another rule which can be computed more easily for large sets. Next, we apply a commonly used concept of the proportional allocation rule $\varphi^{p,\lambda}$, which depends on individual customer arrival rates $\lambda_i$

for the game $(N, C_{opt})$,

$$\varphi_i^{p,\lambda} = \frac{\lambda_i}{\lambda_N} C_{opt}(N) = \frac{h\lambda_i \mu_N^*}{\lambda_N} + \frac{d\lambda_i \lambda_N (1+f)}{2\mu_N^*(\mu_N^* - \lambda_N)}. \tag{2.22}$$

The $\varphi^{p,\lambda}$ rule is reasonable and easy to understand: the coalition cost is paid per demand, and the distributed cost for each provider according to its own customer amount in the pooling system. One of the problems incurred by $\varphi^{p,\lambda}$, is that it may not justify the self-interest of sub-coalitions. However, we prove that this cost allocation rule is stable for the game $(N, C_{opt})$. At the beginning of the proof, we exploit the relationship between $C^*$ and $\lambda$. We denote the unit demand cost by $C_d$, for the total cost $C(\cdot)$ under fixed service capacity situation. It represents the overall cost for one demand per unit time,

$$C_d(\lambda, \mu) = \frac{C(\lambda, \mu)}{\lambda} = c_h \rho^{-1} + \frac{c_w(1+f)}{2} \frac{1}{\rho^{-1}(\rho^{-1} - 1)} \lambda^{-1} = C_d(\lambda, \rho). \tag{2.23}$$

Similarly, the corresponding unit demand cost with optimized service capacity is $C_d^*(\lambda) = C^*(\lambda)/\lambda = C_d(\lambda, \mu^*)$. The following lemma shows the impact of $\lambda$ on $C_d^*$.

**Lemma 2.3.** The unit demand cost $C_d^*$ is decreasing in the customer arrival rate $\lambda$.

*Proof.* Because of Equation (2.23), it is obvious to state that $C_d(\lambda, \rho)$ is decreasing in $\lambda$, for a given $\rho$. Consider two service systems with $\lambda_1 \geq \lambda_2$, and the optimal server utilizations $\rho_1^*$ and $\rho_2^*$, respectively. We may write $C_d^*(\lambda_1) = C^*(\lambda_1)/\lambda_1 = min\{\ C(\lambda_1)/\lambda_1\} = C(\lambda_1, \rho_1^*)/\lambda_1 \leq C(\lambda_2, \rho_1^*)/\lambda_2 \leq C^*(\lambda_2)/\lambda_2 = C_d^*(\lambda_2)$. The first inequality can be concluded from the decreasing of $C_d$ in $\lambda$, and the second one follows from the definition of $C^*(\cdot)$. This finishes the proof of the lemma. □

Consider now Definition 2.4 and 2.5 to define the collection PMAS (Population Monotonic Allocation Scheme). It has been proposed by Sprumont in 1990 [Yves, 1990]. It is the collection of the dynamic allocation rules for TU-games.

**Definition 2.4.** The *population allocation scheme* associated with an allocation rule $\varphi$ defined for a game $(N, C)$, is a collection of the allocations $\varphi_U$ defined by $\varphi$ for all subgame $(U, C)$, $\emptyset \subseteq U \subset N$.

**Definition 2.5.** A *population allocation scheme* is a *population monotonic allocation scheme* (P-MAS), if the allocation $\varphi_{i,U}$ for any player $i$ with any permutation of players is monotonous in $U$.

**Proposition 2.4.** The proportional allocation $\varphi^{p,\lambda}$ is a core allocation for the service pooling game $(N, C_{opt})$. Moreover, the relevant proportional allocation scheme $\varphi_U^{p,\lambda}$, for any $U \subseteq N$, is a PMAS of the game $(N, C_{opt})$.

*Proof.* For the core allocation proof, it is sufficient to prove the individual rationality, the (Pareto-) efficiency and the collective rationality (stand-alone requirement) of $\varphi^{p,\lambda}$. The efficient property has already been taken into account in the definition of $\varphi_i^{p,\lambda}$ in Equation (2.22). From Lemma 2.3, we could conclude that for any $i \in N$ or $U \subseteq N$, we have $C_d^*(\lambda_N) \leq C_d^*(\lambda_i)$ and $C_d^*(\lambda_N) \leq C_d^*(\lambda_U)$. From the definition in Equation (2.23), we have $\varphi_i^{p,\lambda} \leq C_{opt}(\{i\})$ and $\sum_{i \in U} \varphi_i^{p,\lambda} \leq C_{opt}(U)$. This means that the cost allocation $\varphi^{p,\lambda}$ satisfies the individual rationality and the collective rationality. Thus, $\varphi^{p,\lambda}$ is in the core of the game $(N, C_{opt})$. Furthermore, it is straightforward to see from Lemma 2.3 that the relevant allocation scheme $\varphi_U^{p,\lambda}$ is a PMAS of the game $(N, C_{opt})$. For any subset $S \subseteq N$ and any player $j \notin U$, $j \in N$, we have $\varphi_{i,U}^{p,\lambda} = \lambda_i C_{opt}(U)/\lambda_U \geq \lambda_i C_{opt}(U \cup \{j\})/\lambda_{U\cup\{j\}} = \varphi_{i,U\cup\{j\}}^{p,\lambda}$. This completes the proof of the proposition. □

The proportional allocation method $\varphi^{p,\lambda}$ provides therefore a dynamic stable solution to the game $(N, C_{opt})$ and all its sub-games $(U, C_{opt})$. Each independent service provider could reduce its expenses, when a new player joins the coalition. $\varphi^{p,\lambda}$ assigns a positive cost to each participant, and this is not always true for the Shapley value. In fact, players also do care about others' profits. In a cost game, the negative allocation means that some players gain money from a paying activity. This may lead to few unhappy players. Furthermore, only one system should be optimized here. Therefore, $\varphi^{p,\lambda}$ is easier to calculate and to understand than the Shapley value, especially for the case of large number of participants. Without considering individual contributions, this method may have a non-negligible trouble of unfairness. In order to choose an adequate cost-sharing rule for the game $(N, C_{opt})$, we next analytically compare between the two methods.

### 2.5.4. Comparison between $\varphi^{p,\lambda}$ and $sh^{opt}$

We compare the two allocation rules above for our game $(N, C_{opt})$ using the four desirable properties of the Shapley value. It is obvious to see that the dummy player signifies the player which has an empty arrival rate. $\varphi^{p,\lambda}$ is calculated by the single variable $\lambda_i$ and the total cost $C_{opt}(N)$, i.e., $\lambda_i$ is the unique parameter for all players. Thus, the anonymity, efficiency and dummy player property are suitable for the allocation rule $\varphi^{p,\lambda}$ for the game $(N, C_{opt})$. We may then conclude that $\varphi^{p,\lambda}$ is fair for the dummy players and the players with equal arrival rates as $sh^{opt}$.

Another fairness issue is related to the costs distributed for the players with different arrival rates. The game $(N, C_{opt})$ is a single-attribute game of $\lambda_i$. We assume that the company size is directly related to $\lambda_i$. Consider next a situation with three companies, $\{c_h, c_w, f\} = \{1, 3, 0.2\}$ and the customer arrival rates are 1, 1 and 10. Under the $\varphi^{p,\lambda}$ cost allocation, the two small companies with a low arrival rate $\lambda_i = 1$ save much more than the large company ($1.21 > 0.55$). The opposite is true for $sh^{opt}$ with $0.90 < 1.18$. If the large company is not satisfied its saving and leaves the coalition, $\{1, 2\}$ will save much less ($0.46 < 1.21$). The Shapley value is therefore preferred for this small set. Therefore, we look for a fairness condition for choosing between the two rules.

From Lemma 2.3, we conclude that the cost function $C_{opt}$ is elastic as defined in [Özen et al., 2011]. Now, consider a fair property defined for the games with the elastic single-attribute situation [Karsten et al., 2015a].

**Definition 2.6.** For an elastic single-attribute situation, $(N, C, \lambda)$, an allocation rule $\varphi$ is considered to have the *benefit ordering property* (BO), if for all $i, j \in N$ with $\lambda_i \leq \lambda_j$, $C(\{i\}) - \varphi_i \leq C(\{j\}) - \varphi_j$.

This is to say that the saving of each player $C(\{i\}) - \varphi_i$ is non-decreasing in the attribute $\lambda$. It means that the relative large company should save more than the small one. Using Theorem 6.4 in [Karsten et al., 2015a], $sh^{opt}$ follows BO for any concave game. But BO is not always established for $\varphi^{p,\lambda}$. Thus, $sh^{opt}$ provides a more fair allocation rule than $\varphi^{p,\lambda}$ for the players with different $\lambda_i$. We define the saving difference of two players

by $\Delta_{ij} = (C(\{i\}) - \varphi_i) - (C(\{j\}) - \varphi_j)$. We also define that an allocation $\boldsymbol{\varphi}$ is said to be unfair for players $i$ and $j$ in the situation $(N, C, \lambda)$, if $\Delta_{ij}$ is positive for $\lambda_i < \lambda_j$. We prove that this unfairness weakness of $\varphi^{p,\lambda}$ would be reduced in the following situation.

**Lemma 2.4.** A saving difference using the allocation rule $\varphi^{p,\lambda}$ in the game $(N, C_{opt})$, decreases in the number of service providers $n$.

*Proof.* We suppose $\Delta_{12}^{p,\lambda} > 0$ for a pair of players $\{1,2\} = N_0$ with $\lambda_1 < \lambda_2$. We define $|N_{k+1}| = |N_k| + 1, k \in \mathbf{N}^+$ with $\lambda_{N_{k+1}}$ by $\lambda^{k+1} = \lambda^k \cup \{\lambda_{|N_{k+1}|}\}$. Now, let us judge the BO property of $\varphi^{p,\lambda}$ for our game $(N, C^{opt})$ using Definition 2.6. From Lemma 2.3, we have $C_d^*(N_{k+1}) \leq C_d^*(N_k)$. For the allocation rule $\varphi^{p,\lambda}$, we may write.

$$\begin{aligned}\Delta_{12}^{p,\lambda}(N_k) =& (C^*(\lambda_1) - C_d^*(N_k) * \lambda_1) - (C^*(\lambda_2) - C_d^*(N_k) * \lambda_2) \\ =& C_d^*(N_k) * (\lambda_2 - \lambda_1) + (C^*(\lambda_1) - C^*(\lambda_2)) \\ \geq& C_d^*(N_{k+1}) * (\lambda_2 - \lambda_1) + (C^*(\lambda_1) - C^*(\lambda_2)) = \Delta_{12}^{p,\lambda}(N_{k+1}).\end{aligned}$$

Thus, $\Delta_{ij}^{p,\lambda}$ is decreasing in $n$ if $\lambda_i < \lambda_j$. □

We could conclude that the unfairness $\Delta_{ij} > 0$ for $\lambda_i < \lambda_j$ using $\varphi^{p,\lambda}$ could be reduced with new joining players. It is also easy to see that if $\Delta_{12}^{p,\lambda} < 0$ for $\{1,2\} = N_0$ with $\lambda_1 \leq \lambda_2$, BO for these two players is always ensured by $\varphi^{p,\lambda}$ in the game $(N, C_{opt})$ with same parameters $\{c_h, c_w, f\}$. From (ii) in Lemma 2.2, we know that $C^*(\lambda_j) - C^*(\lambda_i)$ is always positive. We therefore may state the following result.

**Lemma 2.5.** For a service pooling game $(N, C_{opt})$, $\varphi^{p,\lambda}$ satisfies the benefit ordering property if

$$C_d^*(\lambda_N) < min\{\frac{C^*(\lambda_j) - C^*(\lambda_i)}{\lambda_j - \lambda_i} | i, j \in N \text{ with } \lambda_i < \lambda_j\}. \tag{2.24}$$

As stated above, we could use $\Phi_{ij} = [C^*(\lambda_j) - C^*(\lambda_i)]/(\lambda_j - \lambda_i) \geq 0$ to judge the saving difference between players $i$ and $j$. We now discuss about the impact of $\lambda_j - \lambda_i$ in the following lemmas.

**Lemma 2.6.** For a service pooling game $(N, C_{opt})$, let $\lambda_{min} = min\{\lambda_i \in N\}$ and $\lambda_{max} = max\{\lambda_j \in N\}$. Then

$$min\{\frac{C^*(\lambda_j) - C^*(\lambda_i)}{\lambda_j - \lambda_i} | i, j \in N \text{ with } \lambda_i < \lambda_j\} = \frac{C^*(\lambda_{max}) - C^*(\lambda_{min})}{\lambda_{max} - \lambda_{min}}. \tag{2.25}$$

*Proof.* We use the monotonicity of $\Phi_{ij}$. First, we fix $\lambda_i$ and change $\lambda_j$ to $\lambda_j + \Delta_\lambda$, which is an infinitesimal of $\lambda$. From the concavity of $C^*$ in $\lambda$ (Theorem 2.4), we have $\partial C^*(\lambda_i)/\partial\lambda \geq \partial C^*(\lambda_j)/\partial\lambda \geq \partial C^*(\lambda_j + \Delta_\lambda)/\partial\lambda$. We have

$$\partial C^*(\lambda_i)/\partial\lambda \geq \frac{C^*(\lambda_j) - C^*(\lambda_i)}{\lambda_j - \lambda_i} \geq \partial C^*(\lambda_j)/\partial\lambda.$$

Then, we may write

$$\frac{C^*(\lambda_j) - C^*(\lambda_i)}{\lambda_j - \lambda_i} \geq \frac{C^*(\lambda_j + \Delta_\lambda) - C^*(\lambda_i)}{\lambda_j + \Delta_\lambda - \lambda_i} \geq \partial C^*(\lambda_j + \Delta_\lambda)/\partial\lambda,$$

which decreases in $\lambda_j$. Using symmetric arguments, we may prove that $[C^*(\lambda_j) - C^*(\lambda_i)]/(\lambda_j - \lambda_i)$ increases in $\lambda_i$. This finishes the proof of the lemma. □

From Lemmas 2.5 and 2.6, we use that the maximum gap of individual arrival rates determines the minimum $\Phi_{ij}$ of a game $(N, C_{opt})$. This leads to the result in the next proposition.

**Proposition 2.5.** For a service pooling game $(N, C^{opt})$, $\varphi^{p,\lambda}$ meets the benefit ordering property if

$$C_d^*(\lambda_N) < \frac{C^*(\lambda_{max}) - C^*(\lambda_{min})}{\lambda_{max} - \lambda_{min}}. \tag{2.26}$$

As stated in Lemma 2.4, $\Delta_{ij}$ is decreasing in $n$, which implies that the left side of Equation (2.26) decreases in $n$. From the proof of Lemma 2.5, the right side of Equation (2.26) is decreasing in $\lambda_{max}$ and increasing in $\lambda_{min}$. Therefore, we could state that $\varphi^{p,\lambda}$ satisfies BO for our game $(N, C_{opt})$ with a sufficiently large $N$ or similar company sizes,

i.e., small value for $\lambda_{max} - \lambda_{min}$. A summary of the comparison between $\varphi^{p,\lambda}$ and $sh^{opt}$ is given in Table 2.5.

| | | Small set with large $\lambda_{max} - \lambda_{min}$ | | Large set with small $\lambda_{max} - \lambda_{min}$ | |
|---|---|---|---|---|---|
| | Stability | Complexity | Fairness | Complexity | Fairness |
| $\varphi^{p,\lambda}$ | + | ++ | - | ++ | + |
| $sh^{opt}$ | + | + | ++ | - | ++ |

**Table 2.5:** *Comparison between $\varphi^{p,\lambda}$ and $sh^{opt}$*

Given the computing complexity of $sh^{opt}$, $\varphi^{p,\lambda}$ is then considered as a suitable cost-sharing method for the game $(N, C_{opt})$, for a large enough set of companies with similar sizes.

**Impact of service time variability $f$.** Unfortunately, it is difficult to analytically evaluate the impact of $f$ on the fairness of $\varphi^{p,\lambda}$ for our game. Consider the unit demand cost $C_d^*$ in $f$ with different $\lambda$ in Figure 2.6.

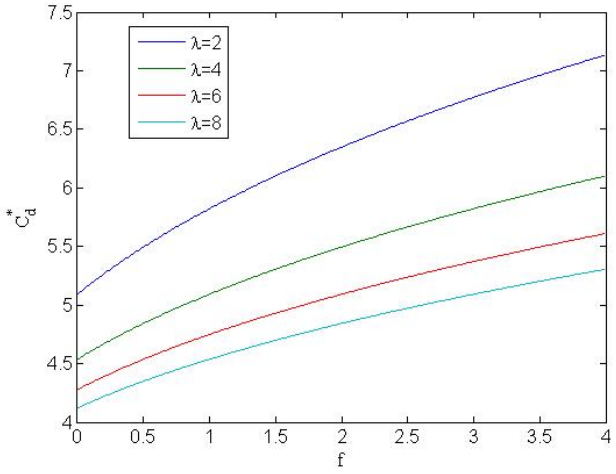



**Figure 2.6:** *$C_d^*$ in $f = (0, 4)$ with different $\lambda = \{2, 4, 6, 8\}$*

It is shown that $C_d^*$ is increasing in $f$ and is more increased with a lower arrival rate $\lambda$. For Equation (2.26), it left side is surely increased in $f$. But it is difficult to analyze its right side. We therefore resort to numerical experiments to investigation this question. We consider five cases of six independent service companies with $c_h = 3$, $c_w = 2$ and

$f = \{0.2, 1\}$, and the customer arrival rate of each company is as shown in Table 2.6. Further numerical examples are given in Appendix A.

| **Company No.** | **1** | **2** | **3** | **4** | **5** | **6** |
|---|---|---|---|---|---|---|
| **Case 1** | 2 | 3 | 4 | 11 | 7 | 10 |
| **Case 2** | 4 | 6 | 8 | 11 | 10 | 2 |
| **Case 3** | 7 | 5 | 11 | 2 | 9 | 3 |
| **Case 4** | 7 | 9 | 12 | 11 | 8.5 | 7 |
| **Case 5** | 2 | 3 | 2.5 | 4 | 3 | 2 |

**Table 2.6:** *5 cases of 6 companies pooling with* $c_h = 3$, $c_w = 2$

Figure 2.7 illustrates the impact of $f$ on the relative difference between $sh^{opt}$ and $\varphi^{p,\lambda}$ for each company. The difference presented in this figure is defined as $(\varphi_i^{p,\lambda} - sh_i^{opt})/\varphi_i^{p,\lambda}$ for a company $i \in \{1, \ldots, 6\}$. We observe that small companies, which have lower customer arrival rates than the average of the coalition, should pay more in $sh^{opt}$ than that in $\varphi^{p,\lambda}$. The opposite is true for large companies. Furthermore, the two allocation rules lead to very near costs for medium companies with arrival rates close to the average. This is because, in the cost allocation $\varphi^{p,\lambda}$, small companies obtain more gains from the unit demand cost reduction than large companies do. Meanwhile, the joining of a small company brings less reduction of $C_d^*$ for the other companies in the coalition, than the joining of a larger one. $sh^{opt}$ accounts for this phenomenon by definition, but not in $\varphi^{p,\lambda}$. Moreover, the system with a higher $f$ obtains more relative cost reduction by collaborating. Thus, the difference increases in $f$.

## 2.6. Conclusion

Using cooperative game theory, we have studied the cost-sharing problem among a set of independent service providers in a complete service capacity pooling system. We extended the existing results to the service pooling game for M/GI/1 service systems. When the service capacities are fixed, the service pooling game would be a sum of two games: an additive service capacity game and a subadditive service congestion game.

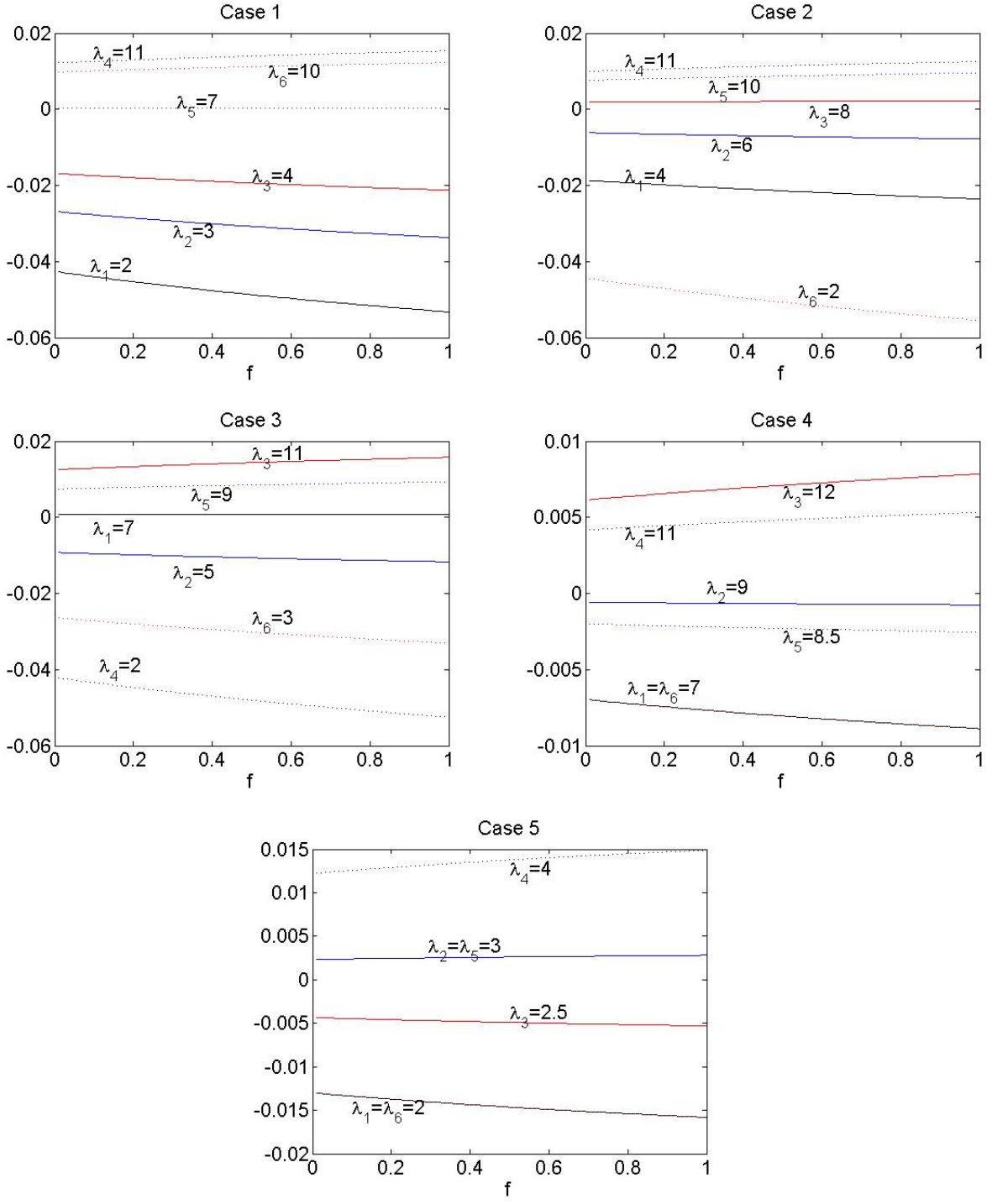



**Figure 2.7:** *Difference between the two allocations with $f$ varying on* $[0,1]$

We proved that a stable cost allocation always exist. Thus, a stable cost-sharing solution could be derived using a mathematical programming approach. When service capaci-

ties are optimized to minimize the total operating cost, we have analyzed the properties related to the optimal service rate. We presented two special stable allocation rules: the well-known Shapley value, and the general proportional allocation rule depending on the individual customer arrival rates. They are both stable solutions for our game $(N, C_{opt})$. Analytical evaluations show that the proportional allocation rule gives a simple cost allocation solution with a good fairness performance in the case of a large set of companies with similar sizes. In the next chapter, we will consider the impact of customer abandonment in service capacity pooling strategy.

# Chapter 3

# Cooperation in Service Systems with Impatient Customers

In this chapter, we consider a group of homogeneous and independent single server service providers with impatience customers, where a customer quits the system without service whenever his waiting time in the queue exceeds his patience time threshold. We study collaboration strategies for the capacity pooling between service providers. The advantage of collaboration in the service systems accounting customer abandonment, is not only the sharing of instant idle resources but also the reducing of abandoned customers. We use the cooperative game theory to analyze the profitable collaborative organization and the cost-sharing method. Under Markovian assumptions for inter-arrival, service and patience times, we define a cooperative game with transferable utility and a fixed service capacity for each coalition. We prove that the grand coalition is the most profitable coalition and that the game has a non-empty core. We then

examine the impact of abandonment on the stability of Shapley value. Furthermore, we prove the concavity of the waiting queue length with respect to the abandonment rate, and give a condition under which the Shapley value is situated in the core. We also study the cost-sharing problem of the relative cooperative game with the optimized service capacity, and prove that the proportional allocation rule based on customer arrival rates gives a dynamic stable allocation to all relative sub-games.

## 3.1. Introduction

In this chapter, we study service capacity pooling strategies under the context of customer impatience (abandonment). A customer who abandons is a customer who decides to leave and give up service if his waiting time in the queue exceeds some random patience threshold. We use also the cooperative game theory to analyze beneficial strategies, their stability issues and how these are affected by customer impatience.

Customer abandonment is an important feature for various practical situations of service systems such as healthcare systems, call centers, telecommunication networks, just to name a few [Jouini, 2012]. Patients waiting for organ transplantation may face a risk of complication or death, which could be modeled as abandonment. A customer who calls a call center is in general willing to wait only a limited amount of time for service to begin. If service has not begun by this time, the customer abandons the queue and is considered as lost [Mandelbaum and Zeltyn, 2009, Jouini et al., 2013, Wallace and Whitt, 2005]. Further examples include visitors that may lose interest for the attractions of a congested amusement park [Kostami and Ward, 2009], passengers who abandon a given type of congested transportation [Shi and Lian, 2016], and post-triage patients who leave without being seen by a physician in an emergency department [Batt and Terwiesch, 2015].

Although its prevalence in practice, published papers related to the analysis of service pooling games, in the presence of impatience, are scarce. Moreover, the queueing literature has shown the importance of incorporating abandonments in order to obtain accurate results [Garnett et al., 2002, Whitt, 2006, Mandelbaum and Zeltyn, 2009]. Existing queueing games mainly deal with models with infinitely patient customers [González and Herrero, 2004, Anily and Haviv, 2010, Garcia-Sanz et al., 2008, Yu et al., 2015, Karsten et al., 2015b]. It is therefore obvious that extending existing studies in the context of abandonment is of value. Our focus here is on the horizontal form of cooperation, i.e., pooling strategies among homogeneous servers. Pooling allows to reduce the number of abandonments through the sharing of idle resources. The pooling advantage for an

alliance is then apparent, but the collective interests cannot be the incentive for each individual service provider to join the coalition. The following questions need then to be addressed: 1) which coalition structure is the most profitable one for the whole group, and 2) how to allocate the total cost among the participants [Cruijssen et al., 2007].

In this chapter, we consider a group of independent single server service providers. Each provider faces its own incoming stream of impatient customers. We suppose that each incoming customer stream is strictly oriented to the corresponding server. Therefore, there is no need to consider competition in the group. A provider could join a profitable coalition, by sharing its capacity, based only on its own benefit from the coalition. To analyze individual and coalition service models, we use cooperative game theory. To the best of our knowledge this is the first contribution to the cooperative literature that accounts for customer impatience. We show that the service pooling strategy among independent service systems, modeled as $M/M/1+M$ queues, is profitable. We prove the non-emptiness of the core for the pooling game under the case of a fixed service capacity. When participants have identical sever loads, we prove that the Shapley value is situated in the core. To prove the result, we first show that the stationary expected queue length and the expected number of customers in the system are both decreasing and convex in the customer abandonment rate (if it is lower than the service rate). This monotonicity result could be also useful for the optimization of queueing systems in general. We then investigate the impact of abandonment on the cost allocation stability of the Shapley value. In the case of optimized service capacities, we prove that the proportional allocation rule, based on individual customer arrival rates, provides a dynamic stable allocation rule to all relative sub-games.

The remainder of this chapter is divided into three sections. In Section 3.2, we describe the individual and collaborative queueing models. In Section 3.3, we define and analyze the service pooling problem with a fixed service capacity. Then, we extend this game to the optimized service capacity case and give a simple formatted core allocation in Section 3.4.

## 3.2. Modeling and observations

### 3.2.1. Service systems modeling with impatience

We consider a group of $n \in \mathbb{N}^+$ independent service providers, denoted by $N = \{1, ..., n\}$. Each service provider has its own single server queue handling its own class of customers in the same queue. For a provider $i$ (system $i$, $i \in N$), we assume that customer arrivals follow a Poisson process with mean rate $\lambda_i$, and service times are independent exponentially distributed random variables with mean $\mu_i^{-1}$. Customer patience times, i.e., the maximum waiting times of customers in the queue, are independent and exponentially distributed with mean $\theta_i^{-1}$. The waiting space is assumed to be large enough such that no customer leaves the system immediately upon arrival due to the waiting space limit. We denote, for system $i$, the expected stationary queue length by $L_q(\lambda_i, \mu_i, \theta_i)$, and the stationary abandonment probability by $P_a(\lambda_i, \mu_i, \theta_i)$. For system $i$, customers are served in the order of their arrivals, i.e., under the first come first serve (FCFS) discipline of service. Following the above assumptions, an individual service provider can be seen as an M/M/1+M queueing model as shown in Figure 3.1.

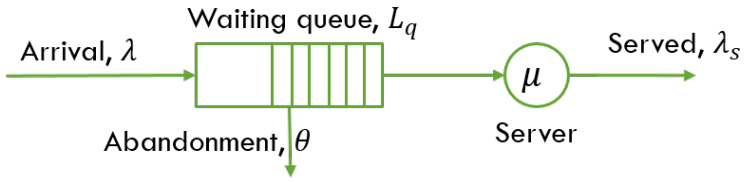



**Figure 3.1:** *An individual service provider, an M/M/1+M queue*

Following the classical assumption as in [González and Herrero, 2004, Anily and Haviv, 2010, Yu et al., 2015], we assume that the service providers operate their service capacities together and run as a 'super-server', i.e., a single server with a high service capacity. For a coalition $U = \{1, \ldots, u\} \subseteq N$, the combined mean arrival rate is $\lambda_U = \sum_{i \in U} \lambda_i$. Since individual arrival streams are independent and follow each a Poisson process with mean rate $\lambda_i$, the resulting stream of arrivals follows also a Poisson process

with mean rate $\lambda_U$. Patience times in a coalition are assumed to be statistically identical to those in individual systems, i.e., exponentially distributed with rate $\theta$. The service times of the pooled system are also exponentially distributed with a mean rate denoted by $\mu_U$. Therefore, the $u$ providers in the coalition $U$ are combined into a new M/M/1+M queueing system.

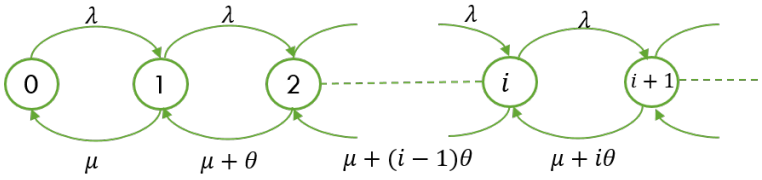



**Figure 3.2:** *Markov chain for the M/M/1+M queue*

Consider next an M/M/1+M queue with arrival rate $\lambda$, service rate $\mu$, and abandonment rate $\theta$. Using the Markovian assumption of patience times, we have $P_a = \theta L_q/\lambda$. To compute $L_q$, we consider the stochastic process describing the number of customers in system. It is a Markov chain as shown in Figure 3.2. Under the stationary regime, we may write

$$L_q = \sum_{j=0}^{\infty}(j-1)p_j = \sum_{j=0}^{\infty}(j-1)\frac{a_j}{\sum_{l=0}^{\infty}a_l}, \tag{3.1}$$

where

$$a_j = \frac{\lambda^j}{\prod_{l=0}^{j-1}(\mu + l\theta)}, \text{ for } j \in \mathbb{N}^+, \text{ and } a_0 = 1, \tag{3.2}$$

and $p_j$ are the stationary system state probabilities, with $p_j = a_j p_0$ for $j \in \mathbb{N}^+$, $p_0 = (\sum_{l=0}^{\infty} a_l)^{-1}$. The effective improvement of the system performances by capacity pooling is shown well in following example.

**Example.** Consider two counters $\{1,2\}$ in a fast food restaurant. Because the two service employees are different skilled, the service rates $\mu_i$ and the customer arrival rates $\lambda_i$, $i = \{1,2\}$ are different for the two counters. Assume that the mean service times $\mu_i^{-1}$ are 5 mins and 3 mins, perspectively. The second counter is much faster than the first one, so that customers prefer the second queue. The mean times inter two continuous arrivals of customer are $\lambda_1^{-1} = 6$ mins and $\lambda_2^{-1} = 4.2$ mins. If a customer waits more

than 10 mins, he would leave without buying any food. When the two counters works individually, we have $L_{q,1} = 0.4430$ and $L_{q,2} = 0.4498$. There are about 11 customers leaving directly without service in two hours.

If the two employees work together for one counter, then it is reasonable to assume that the mean service time is smaller than the faster counter. Assume that $\mu_{\{1,2\}}^{-1} = 2$ mins. For the super-counter $\{1,2\}$, we have $L_{q,\{1,2\}} = 0.7438$ and about 9 lost customers in two hours. There are less customers waiting in the queue and less customers losing per hour before service.

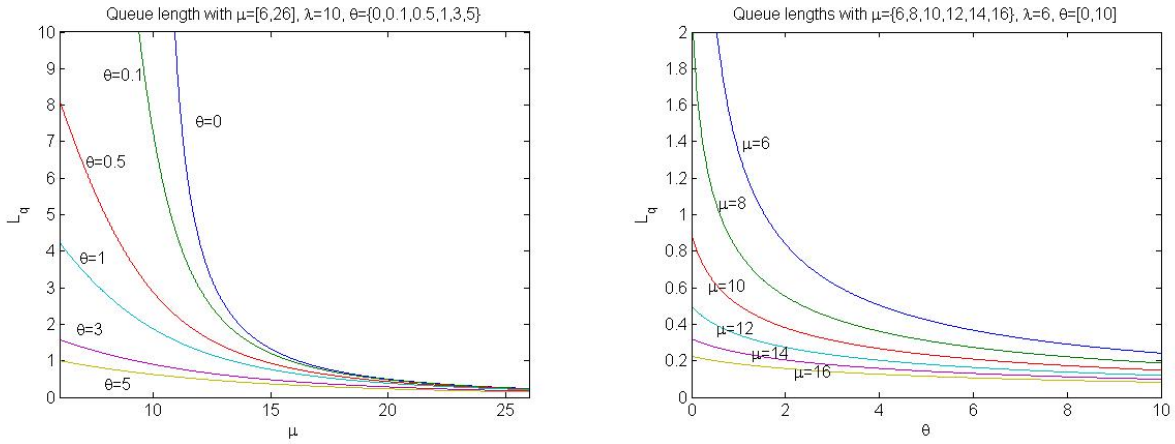
To define the total system cost in this M/M/1+M coalition queue, we consider three types of linear costs: the capacity holding cost $C_h$ which may consist of equipment maintenance fees and staff salary; the customer waiting cost in the queue $C_w$; and the customer abandonment cost $C_a$. All cost components are defined per time unit and the corresponding cost parameters are denoted $c_h$, $c_w$ and $c_a$, respectively. Thus, the total system operating cost, say $C_{total}$, is

$$\begin{aligned} C_{total}(\lambda, \mu, \theta) &= C_h + C_w + C_a \\ &= c_h\mu + c_w L_q + c_a \lambda P_a \\ &= c_h\mu + (c_w + c_a\theta)L_q. \end{aligned} \tag{3.3}$$

### 3.2.2. Observation of queue length and abandonment probability

In queueing systems designing, it is interesting to determine whether an objective function or a performance measure is convex/concave or not. We observe here the convexity and the monotonicity of $L_q$ and $P_a$ in $\mu$ and $\theta$, which is illustrated in Figure 3.3 and 3.4.

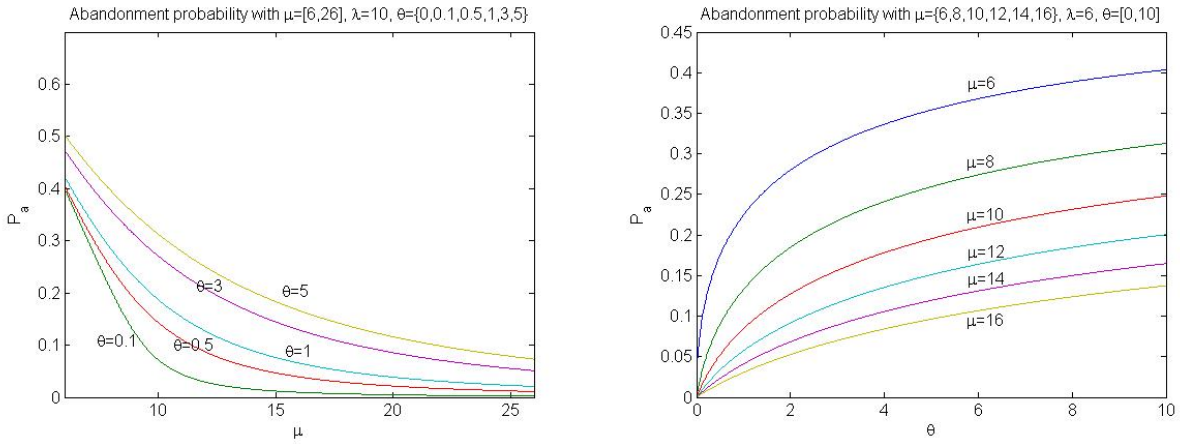
From Figure 3.3(a), we find that the queue lengths $L_q$ with different abandon rate $\theta = \{0, 0.1, 0.5, 1, 3, 5\}$ close together with the augmentation of the service rate $\mu$. This is because that if the service rate $\mu$ is sufficient larger than the arrival rate $\lambda$ and the abandon rate $\theta$, majority of waiting customers are serviced before their tolerant waiting times, and the queue length $L_q$ is reduced toward to 0. The analogous phenomenon




(a) (b)

**Figure 3.3:** *Queue length $L_q$ in $\mu$ and $\theta$*

appears in the augmentation of $\theta$ with different $\mu = \{6, 8, 10, 12, 14, 16\}$ in Figure 3.3(b). When $\theta >> \mu$, the overwhelming majority of the waiting customers abandons before service begins and the service system is close to an M/M/1/1 system.




(a) (b)

**Figure 3.4:** *Abandonment probability $P_a$ in $\mu$ and $\theta$*

In Figure 3.4(a), we observe that the abandon probabilities $P_a$ with different $\theta$ diverge with the augmentation of $\mu$ at the beginning and converge at infinity (almost all customers received services). In Figure 3.4(b), contrary to $\mu$, we find that $P_a$s are increasing in $\theta$ and divergent at infinity.

From Figures 3.3 and 3.4, it is shown that the service systems with a higher utilisation $\rho = \lambda/\mu$ are more sensitive to the abandonment rate $\theta$, and the service systems with a high enough $\theta$ are less sensitive to the $\rho$. It is obvious that the overall cost defined in Equation (3.3) is sensitive to $\theta$. The relative costs of different systems are sensitive to $\theta$ in the interval $(0, a)$ with $\mid L_q(\lambda, \mu, a) - L_{q,M\backslash M\backslash 1\backslash 1}(\lambda, \mu) \mid \leq \epsilon$.

## 3.3. Collaboration under a fixed service capacity

In this section, we assume that the service capacity of the pooled server is the sum of those of the individual servers, $\mu_U = \sum_{i \in U} \mu_i$ for a coalition $U \subseteq N$. This corresponds to situations where changing the equipment or the physical location is too expensive or almost impossible. Thus, we obtain the following total cost, for a coalition $U \subseteq N$,

$$C_{fix}(U) = C_{total}(\lambda_U, \mu_U, \theta) = c_h \cdot \mu_U + (c_w + c_a\theta)L_q(\lambda_U, \mu_U, \theta), \\ \text{with } \lambda_U = \sum_{i \in U} \lambda_i, \text{ and } \mu_U = \sum_{i \in U} \mu_i. \tag{3.4}$$

For a nonempty finite set $N = \{1, ..., n\}$, we assume that every player could only participate in one coalition $U \subseteq N$, and the total cost of the coalition could be shared among its members with no constraint. Therefore, we can define a cost TU-game associated with the characteristic function (c.f.) of the total cost for a set $N$ of service providers. Recall that the abandonment rate and the cost parameters are identical for all providers in the group $N$. Since there are only two individual parameters $\lambda_i$ and $\mu_i$ for each provider $i$, it suffices to consider the simplified cost expression given by

$$C_{fix}(\emptyset) = 0, C_{fix}(U) = (c_w + c_a\theta)L_q(\lambda_U, \mu_U), \text{ for any } U \text{ with } U \subseteq N, \tag{3.5}$$

as c.f. and the game as $(N, C_{fix})$. In what follows, the analysis of the game $(N, C_{fix})$ is separated into two parts. First, we investigate the existence of stable cost allocations. We then analyze the impact of abandonment on the stability of the Shapley value.

### 3.3.1. Non-emptiness of the core of the game $(N, C_{fix})$

We first show that resource pooling in the context of customer impatience always leads to a total cost reduction for the coalition. Proposition 3.1 shows the advantage of pooling on the service quality in terms of $L_q$.

**Proposition 3.1.** The expected queue length $L_q$ of the pooled server with a fixed capacity is subadditive.

*Proof.* Consider two M/M/1+M systems 1 and 2 with the parameters $\lambda_1, \lambda_2, \mu_1, \mu_2$ and a same $\theta$. We have $\lambda_{1,2} = \lambda_1 + \lambda_2$ and $\mu_{1,2} = \mu_1 + \mu_2$ for the pooled system $\{1,2\}$. We denote the number of customers in the queue by the random variable $Y_{(\cdot)}$. The subadditive problem requires to prove that

$$L_{q,\{1,2\}} \leq L_{q,1} + L_{q,2},$$

which is equivalent to

$$\frac{P(Y_{1,2} = k+1)}{P(Y_{1,2} = k)} \leq \frac{P(Y_1 + Y_2 = k+1)}{P(Y_1 + Y_2 = k)}, \text{ for } k \in \mathbb{N},$$

see page 208 in [Ferguson, 2014]. We define $p_{i,j}$ as the probability of having $j$ customers waiting in the queue $i$, $i \in \{1, 2, \{1,2\}\}$. We may write, for $k \in \mathbb{N}$.

$$\begin{aligned}
P(Y_1 + Y_2 = k+1) &= P(Y_1 + Y_2 = k)[P(Y_1 + 1 | Y_1 + Y_2 = k) + P(Y_2 + 1 | Y_1 + Y_2 = k)] \\
&= P(Y_1 = l+1, Y_2 = j | l + j = k) + P(Y_1 = l, Y_2 = j+1 | l + j = k) \\
&= \sum_{l+j=k} p_{1,l+2} p_{2,j+1} + \sum_{l+j=k} p_{1,l+1} p_{2,j+2} \\
&= \sum_{l+j=k} \frac{\lambda_1}{\mu_1 + (l+2)\theta} p_{1,l+1} p_{2,j+1} + \sum_{l+j=k} \frac{\lambda_2}{\mu_2 + (j+2)\theta} p_{1,l+1} p_{2,j+1} \\
&> \frac{\lambda_1}{\mu_1 + (k+2)\theta} \sum_{l+j=k} p_{1,l+1} p_{2,j+1} + \frac{\lambda_2}{\mu_2 + (k+2)\theta} \sum_{l+j=k} p_{1,l+1} p_{2,j+1} \\
&> \frac{\lambda_1 + \lambda_2}{\mu_1 + \mu_2 + (k+2)\theta} P(Y_1 + Y_2 = k)
\end{aligned}$$

$$= \frac{P(Y_{1,2} = k+1)}{P(Y_{1,2} = k)} P(Y_1 + Y_2 = k).$$

Thus, the queue length $L_q$ is subadditive with pooling. This finishes the proof of the proposition. □

From Proposition 3.1, one can see that the overall expected number of customers waiting in the overall queue length would be minimized in the grand coalition $N$. From Equation (3.5), the c.f. of our game is proportional to $L_q$. Therefore, $(N, C_{fix})$ is subadditive and $N$ is the most profitable coalition structure. We next focus on the stability of the coalitions, that need to prove the existence of stable cost-sharing allocations for the game. For the analytical tractability, we examine the cases with $\mu \geq \theta$. This assumption may be reasonable in practice. It is expected that customers would be willing to wait longer than their service times in average [Mandelbaum and Zeltyn, 2009]. Before giving one of the main results in Theorem 3.1, we provide some required monotonicity results in Lemmas 3.1 and 3.2.

**Lemma 3.1.** The queue length $L_q$ in an M/M/1+M queue is decreasing in $\mu \in \mathbb{R}^+$; and convex in $\mu$ for $\mu \geq \theta$.

*Proof.* Consider two M/M/1+M systems, denoted by systems 1 and 2 with different service rates $\mu_1 \leq \mu_2$ and same $\lambda$ and $\theta$. Let $L_{q,1}$ and $L_{q,2}$ denote the waiting queue lengths for systems 1 and 2, respectively. First, we prove that $L_q$ decreases in $\mu$. Let $Y_{(\cdot)}$ be the random variable measuring the number of waiting customers in the queue. We have

$$\frac{\lambda}{\mu_1 + (k+1)\theta} \geq \frac{\lambda}{\mu_2 + (k+1)\theta}.$$

Then $P(Y_1 = k+1)/P(Y_1 = k) \geq P(Y_2 = k+1)/P(Y_2 = k)$, which implies $\{Y_1(t)\} \geq_{st} \{Y_2(t)\}$. This is equivalent to $\mathbb{E}(Y_1) \geq \mathbb{E}(Y_2)$, then $L_{q,1} \geq L_{q,2}$.

To prove the convexity result, we denote the numbers of customers in systems by a sequence of random variables $X(\mu) = \{X(m) | m \in \mathbb{N}\}$ in discrete time. Let $X(\mu)^+ = (X(\mu) - 1)^+$ denote the number of customers in the queue, with $f^+ = x$ if $f \geq 0$, and

$x^+ = 0$ otherwise. As shown in Theorem 3 by [Armony et al., 2009], $X$ is stochastically decreasing and convex in sample path sense (denoted by SDCX(sp)) in $\mu$ with $\mu \geq \theta$. Because $f = (x-1)^+$ is increasing and convex, it follows that $X^+$ is also SDCX(sp) by Proposition 3.2(b) in [Shaked and Shanthikumar, 1988]. From Theorem 3.6 in [Shaked and Shanthikumar, 1988], $L_q = \mathbb{E}(X^+)$ is therefore decreasing and convex in $\mu$ with $\mu \geq \theta$. The lemma result follows. □

Note that the convexity result in Lemma 3.1 is also verified for $0 < \mu < \theta$ through an extensive numerical study, although we could not prove it rigorously.

**Lemma 3.2.** The abandonment probability $P_a$ is increasing in the customer abandonment rate $\theta \in [0, +\infty)$.

*Proof.* Let us denote the mean number of customers in service by $L_s$. Let us also denote by $\lambda_a$ and $\lambda_s$ the flows of abandonment and served rates, respectively. In the single server system $L_s = 1 - p_0$. Also, $L_s$ is equal to the server's utilisation $\rho_s = \lambda_s/\mu$ using Little's Law in [Little, 1961]. Thus,

$$\begin{aligned} L_s =& \frac{\lambda - \lambda_a}{\mu} = \frac{\lambda - \theta L_q}{\mu} = 1 - p_0 \\ &\Rightarrow P_a = \frac{\theta L_q}{\lambda} = 1 - \frac{\mu}{\lambda}(1 - p_0), \end{aligned}$$

where $p_0 = (\sum_{j=0}^{\infty} a_j)^{-1}$. It is obvious that $a_j$ as defined in Equation (3.2) is decreasing in $\theta$, for $j \in \mathbb{N}$. Then, $p_0$ is increasing in $\theta$. The abandonment probability $P_a$ is therefore increasing in $\theta$. The proof is completed. □

Using Lemmas 3.1 and 3.2, we obtain the existence of stable cost allocations in Theorem 3.1.

**Theorem 3.1.** The pooling game $(N, C_{fix})$ has a non-empty core for $\mu \geq \theta$.

*Proof.* For any balanced collection B on N, we have

$$C_{fix}(N) = (c_w + c_a\theta)L_q(\lambda_N, \mu_N)$$

$$= (c_w + c_a\theta)L_q(\lambda_N, \sum_{U\in B} \beta_U\mu_U\frac{\lambda_N}{\lambda_U}\cdot\frac{\lambda_U}{\lambda_N}) \tag{3.6}$$

$$\leq (c_w + c_a\theta)\cdot\sum_{U\in B}\beta_U\frac{\lambda_U}{\lambda_N}L_q(\lambda_N, \mu_U\frac{\lambda_N}{\lambda_U}) \tag{3.7}$$

$$= (c_w + c_a\theta)\cdot\sum_{U\in B}\beta_U\frac{\lambda_U}{\lambda_N}L_q(\lambda_N, \mu_U\frac{\lambda_N}{\lambda_U}, \theta)$$

$$= \sum_{U\in B}\beta_U(c_w + c_a\theta)\cdot\frac{\lambda_U}{\lambda_N}L_q(\lambda_U, \mu_U, \theta\frac{\lambda_U}{\lambda_N})$$

$$\leq \sum_{U\in B}\beta_U(c_w + c_a\theta)\cdot L_q(\lambda_U, \mu_U, \theta) \tag{3.8}$$

$$= \sum_{U\in B}\beta_U C_{fix}(U).$$

From the definition of a balanced collection and the additivity of $\mu$, it exists $\beta_U$ with $\sum_{U\in B}\beta_U\mu_U = \mu_N$ to guarantee Equality (3.6). Inequality (3.7) holds by the convexity property of $L_q(\mu)$ in Lemma 3.1 and $\beta_U$ with $\sum_{U\in B}\beta_U\lambda_U = \lambda_N$ from the balanced collection definition and the additivity of $\lambda$. Since $P_a(\theta\frac{\lambda_U}{\lambda_N}) \leq P_a(\theta)$ (proved in Lemma 3.2) with $\theta\frac{\lambda_U}{\lambda_N} \leq \theta$ and same $\lambda_U$, we have $L_q(\theta\frac{\lambda_U}{\lambda_N})\cdot\theta\frac{\lambda_U}{\lambda_N} \leq L_q(\theta)\cdot\theta$ which leads to the inequality in (3.8). Thus, the game $(N, C_{fix})$ is a balanced game. According to "Bondareva-Shapley Theorem" in [Bondareva, 1963], the game $(N, D_{fix})$ has a non-empty core, which completes the proof of the theorem. □

### 3.3.2. Impact of abandonment on the stability of the Shapley value

From Theorem 3.1, we conclude that a core allocation always exists for the game $(N, C_{fix})$. Thus, the explicit numerical solutions could be computed through a mathematical programming method, such as the Nucleolus [Bondareva, 1963], the Equal Profit Method [Frisk et al., 2010] or the EPM based on Contribution Weights [Peng et al., 2015]. To the contrary to the case with no abandonments, we find that the presence of impatience could impact the stability of the Shapley value for our game $(N, C_{fix})$. First of all, we use an example of three players to illustrate this feature.

**Example.** As that shown in last chapter, the example of three players game among

M/GI/1 service systems (relevant initial values in Table 3.1) is not concave and the Shapley value could not provide a stable allocation.

| Players | Parameters | | | |
|---|---|---|---|---|
| | $\lambda_i$ | $\mu_i$ | $C_i$ | $sh_i^{fix}$ |
| 1 | 9 | 10 | 8.1 | 3.1329 |
| 2 | 5 | 10 | 0.5 | $-0.9089$ |
| 3 | 2 | 10 | 0.05 | $-1.6145$ |

**Table 3.1:** *3 players pooling game*

Let us now investigate the impact of abandonment on the stability of the Shapley value, which is symbolized by $sh^{fix}$ in $(N, C_{fix})$. For simplification, we choose $c_w = c_a = 1$ which does not affect the stability of the Shapley value. By the increasing of $\theta$, we numerically observe that $sh^{fix}$ provides a stable allocation for $\theta \geq 0.42$. We denote this value as the lower bound of the interval which has $sh^{fix}$ in the core by $\theta^{low}$. In Table 3.2, we calculate coalition and $sh^{fix}$ distributed costs in this case for $\theta = 0$, $\theta^{low}$ and 1000.

| Coalitions | $C_{fix}(U)$ | $\sum_{i\in U} sh_i^{fix}$ | $C_{fix}(U)$ | $\sum_{i\in U} sh_i^{fix}$ | $C_{fix}(U)$ | $\sum_{i\in U} sh_i^{fix}$ |
|---|---|---|---|---|---|---|
| | $\theta = 0$ | | $\theta = 0.42$ | | $\theta = 1000$ | |
| {1,2,3} | 0.6095 | 0.6095 | 0.7827 | 0.7827 | 5.4647 | 5.4647 |
| {1,2} | 1.6333 | <u>2.2240</u> | 1.7250 | 1.5205 | 5.7038 | 5.4766 |
| {2,3} | 0.1885 | $-2.5234$ | 0.2490 | $-0.7716$ | 1.7902 | 1.5626 |
| {1,3} | 0.6722 | <u>1.5184</u> | 0.8190 | 0.8165 | 3.8576 | <u>3.8902</u> |

**Table 3.2:** *Coalition and distributed costs with* $\theta = 0$, 0.42 *and* 1000

The *excess* of a coalition $U \subseteq N$ at $\mathbf{x}$ is defined as the quantity $e_{x,U} = C(U) - \sum_{i\in U} x_i$, i.e., the cost reduction of $U$ by $\mathbf{x}$, which is used to measure satisfaction of coalitions. If $e_{x,U} \leq 0$, the coalition $U$ would like to split off from $N$ by using allocation $\mathbf{x}$. From the comparison between the first two columns in Table 3.2, we observe two unsatisfied coalitions $\{1,2\}$ and $\{1,3\}$, with $e_{sh^{fix},\{1,2\}} < 0$ and $e_{sh^{fix},\{1,3\}} < 0$ for $\theta = 0$. It is to say that 1 leaving $N$ with 2 or 3 could form a more profitable smaller coalition. However, as

$\theta$ increases, $e_{sh^{fix},\{1,2\}}$ and $e_{sh^{fix},\{1,3\}}$ increase and grow gradually toward positive values as shown in Figure 3.5. It means that the presence of $\theta$ enforces the stability of $sh^{fix}$ in this range.

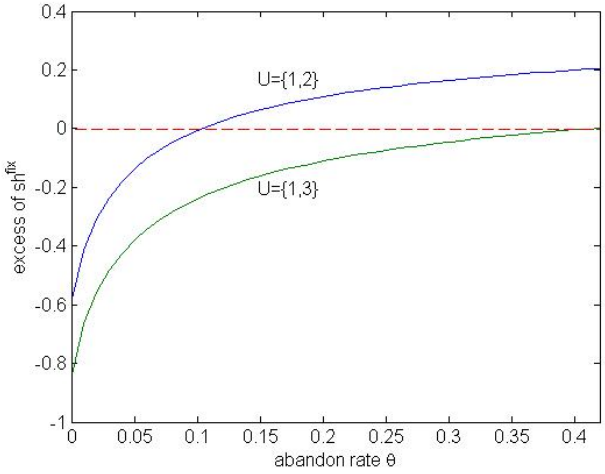



**Figure 3.5:** $e_{sh^{fix},\{1,2\}}$, $e_{sh^{fix},\{1,3\}}$ *in* $\theta$ *for the 3 players game*

By increasing $\theta$, we observe that the coalition $\{1,3\}$ is no longer satisfied by $sh_{fix}$ for $\theta \geq 382.4$, i.e., $e_{sh^{fix},\{1,3\}} < 0$. In a similar way, we denote this point by $\theta^{up}$ as the upper bound of the stable interval of $sh^{fix}$. This interval is denoted by $\Theta = [\theta_{low}, \theta_{up})$. When $\theta$ is much larger than $\lambda$ and $\mu$ ($\theta = 1000$ as given in Table 3.2), the players act as M/M/1/1 queueing systems (no queue). We see that the same coalition $\{1,3\}$ obstruct the grand coalition under $sh^{fix}$. Moreover, as $\theta \to +\infty$, the system abandonment rate converges to $P_a = \lambda/(\lambda+\mu)$ and the expected queue length $L_q$ converges to 0. The total cost function could be then calculated as, $C_{fix} = c_a\lambda^2/(\lambda+\mu)$. This also leads to an unsatisfied coalition $\{1,3\}$ for $\theta = 1000$. More detailed illustrations with higher numbers of players are given in Appendix B.

We observe that the Shapley value may lie at the core with the presence of customer abandonment. Furthermore, we prove that it is in the core, when all providers have identical offered loads ($sh^{fix}$ is stable). Before giving the proof, we first need to analyze the convexity of $L_q$ in $\theta$.

**Theorem 3.2.** The expected queue length $L_q$ of an M/M/1+M system is decreasing in $\theta \in \mathbb{R}^+$; and convex in $\theta$ for $\theta \leq \mu$.

The proof of Theorem 3.2 is given in Appendix C. In this proof, we use sample-path arguments similarly to [Armony et al., 2009], where the authors show a similar result for the relationship between $L_{sys}$ (the expected number of customers in the system) and $\mu$ or the number of servers in Erlang-A model. Note that this convexity property is also useful for various optimization studies in M/M/1+M queueing systems. Now, we are ready to prove the stability of the Shapley value under the offered load constraint.

**Proposition 3.2.** The Shapley value is a core allocation for the game $(N, C_{fix})$, if $\lambda_i/\mu_i = \rho$ for all $i \in N$ and $\theta \leq \mu$.

*Proof.* To prove the stability of the Shapley value, it suffices to prove the concavity of the operational cost $C_{fix}(\rho\mu, \mu, \theta)$ in the service rate $\mu$. Consider any pair of coalitions $U$ and $T$ with $\emptyset \subseteq U \subset T \subset N$ and a service provider $l \in N \setminus T$. We define $\theta_0$ as $\theta/\mu$. Since $C_{fix}$ is twice differentiable in $\mu$, we have

$$\begin{aligned}\frac{\partial^2 C_{fix}(\rho\mu, \mu, \theta)}{\partial \mu^2} &= \frac{\partial^2 C_{fix}(\rho, 1, \theta/\mu)}{\partial \mu^2} \\ &= (c_w + c_a\theta)\frac{\partial^2 Lq(\rho, 1, \theta_0)}{\partial \mu^2} \\ &= (c_w + c_a\theta)[(-\frac{\theta}{\mu^2}) \cdot \frac{\partial^2 Lq}{\partial \theta_0^2} + \frac{2\theta}{\mu^3} \cdot \frac{\partial L_q}{\partial \theta_0}] \leq 0, \end{aligned} \tag{3.9}$$

where the inequality in (3.9) holds from $\partial^2 L_q/\partial\theta^2 \geq 0$ and $\partial L_q/\partial\theta \leq 0$, which is based on Theorem 3.2. Thus, $C_{fix}$ is concave in $\mu$ if $\lambda_i/\mu_i = \rho$. Furthermore, the Shapley value stays in the core as shown in Theorem 7 in [Shapley, 1971]. This finishes the proof of the proposition. □

From Proposition 3.2, we then deduce that the Shapley value is a sufficient reasonable cost allocation for a group of service providers with similar offered loads.

## 3.4. Collaboration under the optimized service capacity

In certain cases, systems could be engineered. For example, the service capacity could be increased according by training employees while may improve the service rate. Assuming here that any individual or coalition chooses the service capacity which minimizes the total system cost, this may lead to less costly coalitions. In this section, we consider a TU-game under the setting with the optimized service capacity. We model the service rate as a continuous variable for each individual or coalition. We denote by $\mu^*$ the optimized service capacity, which is defined by

$$\mu^*(\lambda) = argmin\{c_h\mu + (c_w + c_a\theta)L_q(\lambda, \mu, \theta)|\mu \geq 0\}. \tag{3.10}$$

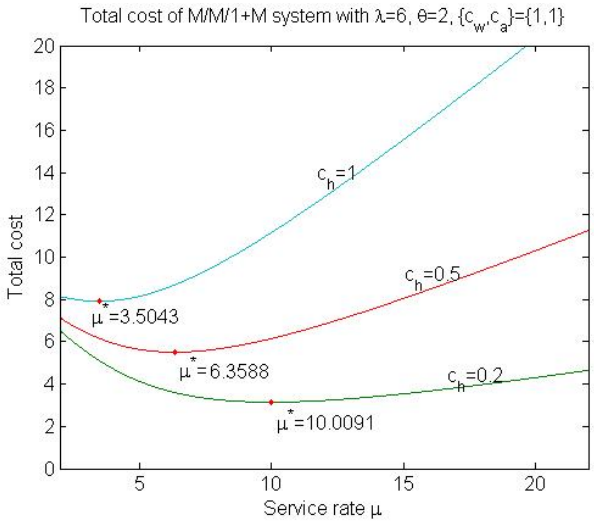
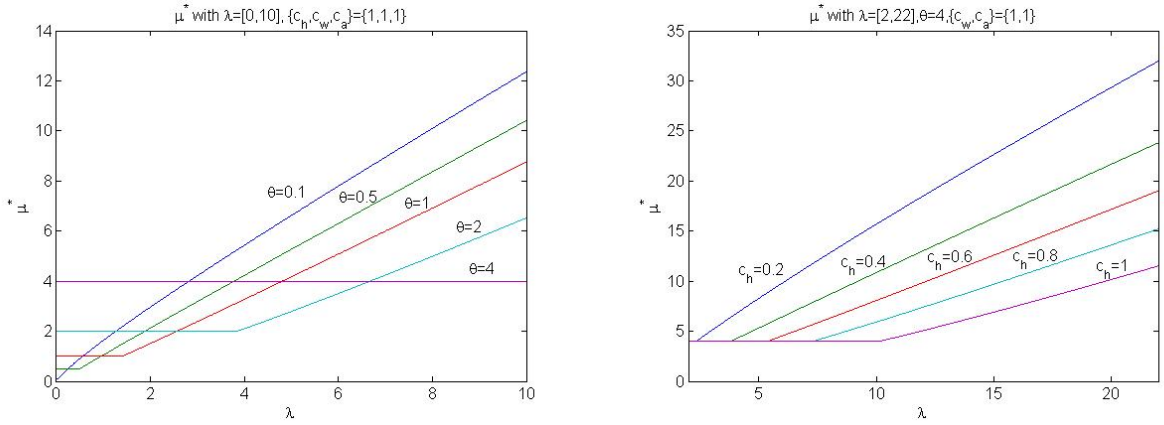



**Figure 3.6:** *Total cost $C$ in $\mu$ with different $c_h$*

**Lemma 3.3.** $\mu^*$ exists and is unique in $[\theta, +\infty)$.

*Proof.* $C = c_h\mu + (c_w + c_a\theta)L_q$ contains a linear function of $\mu$ and a term in $L_q$, which is decreasing in $\mu$ and convex if $\mu \geq \theta$ (Lemma 3.1). Thus, $C$ is positive and convex in $\mu$ if $\mu \geq \theta$. Therefore, $\mu^*$ exists and is unique in $[\theta, +\infty)$. □

From Lemma 3.3, we obtain a unique optimized service capacity on its domain. Using




(a) (b)

**Figure 3.7:** *The optimized service capacity $\mu^*$ in $\lambda$ with different $\theta$ and $c_h$*

its definition in (3.10), it is obvious to see that $\mu^*$ is decreasing in $c_h$. An illustration is given in Figure 3.6. For a coalition $U = \{1, ..., u\} \in N$, we also use the pooled arrival rate $\lambda_U = \sum_{i \in U} \lambda_i$ and an identical abandonment rate $\theta$ as the defined game in Section 3.3. Thus, we could define the optimized cost function using four parameters $\{\theta, c_h, c_a, c_w\}$ and a variable $\lambda_i$ for $i \in U$ as

$$\begin{aligned} C_{opt}(\lambda_U) =& c_h \cdot \mu_U^*(\lambda_U) + (c_w + c_a\theta) L_q(\lambda_U, \mu_U^*(\lambda_U)), \\ & \text{with } \lambda_U = \sum_{i \in U} \lambda_i, \mu_U^* = argmin\{C_{total} | \mu \geq \theta\}, \text{ for any } U \subset N. \end{aligned} \tag{3.11}$$

The pooling game under the optimized service capacity could be then defined as $(N, C_{opt})$. In order to propose a simple stable allocation rule, we next analyze the relationship between the total cost for one demand and $\lambda$.

**Lemma 3.4.** The total cost per unit demand with optimized service capacities $C_{opt}/\lambda$, is decreasing in the customer arrival rate $\lambda$.

*Proof.* Consider two M/M/1+M queues, denoted by 1 and 2, with $\lambda_1 \leq \lambda_2$, $\mu_1^*$, $\mu_2^*$ (associated optimized service capacities) and a same $\theta$. We have

$$\frac{C_{opt}(\{1\})}{\lambda_1} = c_h \frac{\mu_1^*}{\lambda_1} + \frac{(c_w + c_a\theta)}{\lambda_1} L_q(\lambda_1, \mu_1^*, \theta)$$

$$
\begin{aligned}
&= c_h \frac{\mu_1^*}{\lambda_1} + \frac{(c_w + c_a\theta)}{\lambda_1} L_q(\lambda_2, \mu_1^* \cdot \frac{\lambda_2}{\lambda_1}, \theta\frac{\lambda_2}{\lambda_1}) \\
&\geq c_h \frac{\mu_1^*}{\lambda_1} + \frac{(c_w + c_a\theta)}{\lambda_1} \cdot \frac{\lambda_1}{\lambda_2} L_q(\lambda_2, \mu_1^* \cdot \frac{\lambda_2}{\lambda_1}, \theta) \qquad (3.12) \\
&= c_h \mu_1^* \cdot \frac{\lambda_2}{\lambda_1} \cdot \frac{1}{\lambda_2} + \frac{(c_w + c_a\theta)}{\lambda_2} L_q(\lambda_2, \mu_1^* \cdot \frac{\lambda_2}{\lambda_1}, \theta) \\
&= \frac{1}{\lambda_2} C(\lambda_2, \mu_1^* \cdot \frac{\lambda_2}{\lambda_1}, \theta) \geq \frac{C_{opt}(\{2\})}{\lambda_2},
\end{aligned}
$$

where Inequality (3.12) holds by the monotonicity of $P_a$ in $\theta$ in Lemma 3.2. Thus, $C_{opt}/\lambda$ decreases in $\lambda$. This finishes the proof of the lemma. □

From Lemma 3.4, we could conclude that the following allocation rule provides a PMAS for the game $(N, C_{opt})$:

$$\varphi_i^p = \frac{\lambda_i}{\lambda_N} C_{opt}(N), \forall i \in N. \qquad (3.13)$$

**Proposition 3.3.** The propositional allocation $\varphi^p$ defined in Equation (3.13) is a core allocation for the game $(N, C_{opt})$, and the associated allocation scheme $\varphi^{as,p}$ gives a PMAS for this game.

As proved in Section 2.5.2, the game with $\theta = 0$ provides a concave TU-game. We next consider the other limit for $\theta$ much higher than $\lambda$ and $\mu$, i.e., where a provider approaches an M/M/1/1 queue. We rewrite the cost function as

$$C_{opt} = c_h\mu + c_a \frac{\lambda^2}{\lambda + \mu}. \qquad (3.14)$$

With this cost function, we could obtain the optimized service rate as follows.

$$\mu^* = \begin{cases} (\sqrt{c_a/c_h} - 1)\lambda, \text{if } c_a > c_h \\ 0, \text{ otherwise.} \end{cases} \qquad (3.15)$$

When $c_a \leq c_h$, we get $\mu^* = 0$. It means that the optimal choice of the server is not working. In the remaining case, the total cost $C_{opt}$ is linear with $\lambda$ and the game $(N, C_{opt})$ turns to a linear game.

## 3.5. Conclusion

We considered the pooling problem for service systems while accounting for the important feature of customer abandonment. We investigated the cooperative strategy among independent service providers with a complete service capacity pooling. When the service providers directly combine their queues and service capacities, we proved the non-emptiness of the core under the situation with a fixed service capacity. With customer abandonment, the stability of the Shapley value could be affected. We studied the convexity of the expected queue length in the abandonment rate. This result is also helpful when addressing other design issues. Furthermore, we found that if all service providers have an identical offered load, the Shapley value is absolutely stable for our game. Under the optimized service capacity, we proved that a simple proportional allocation rule provides a stable allocation for the relative game. This chapter and Chapter 2 are both assumed in the 'super-server' pooling environment. We will consider the multi-server pooling case in the next chapter.

# Chapter 4

# Collaboration for Multi-server Service Systems

In Chapters 2 and 3, we analyzed resource pooling strategies under 'super-server' assumptions. In this chapter, we study the effect of resource pooling on system performance and profit for multi-server service systems incorporating customer abandonment, with the goal of evaluating whether the results under the 'super-server' assumption are still suitable in the multi-server case.

We consider two single-class queueing systems with different setting of pooling, i.e., using the 'super-server' assumptions and another using identical parallel servers. The pooling strategy efficiency is estimated via the expected number of customers in each system and the mean probability of customer abandonment. Although it is intuitive to expect efficiency improvements in the pooled multi-server system, it is no longer obvious to conclude that all members will benefit from pooling as we proved in previous chapters. We compare between the two

pooling settings from a coalition perspective. We numerically evaluate the effects of service duration variability and customer abandonment on the two corresponding games using the Shapley value and the nucleolus allocations. Furthermore, we show that the service pooling in multi-class systems, is not always profitable because of the additional variability among customer classes.

## 4.1. Introduction

In this chapter, we address the service pooling problem for multi-server pooling systems. We allow that customers to be impatient in each individual or pooled service system. They might decide to leave (abandon) before their services begin when their waiting times are expired. We compare two corresponding TU-games of two typical pooling settings and analyze dependent service pooling problems using cooperative game theory. In the rich literature on capacity pooling research, there are two typical settings for the pooled systems: the 'super-server' setting, i.e., all services resources are pooled in one single server, and the multi-server setting, i.e., all servers work in parallel in a common service factory. Various examples of both setting are discussed in [Kleinrock, 1976]. The two pooling settings have been extensively used to evaluate the overall gains from pooling by comparing the system performance of the pooled system and the individual systems [Mandelbaum and Reiman, 1998, Iyer and Jain, 2004, Tekin et al., 2014, Kim and Kim, 2015, Andradóttir et al., 2017]. The main reason for the 'super-server' assumption is that dealing with multi-server queues with general service times and customer abandonment is very hard, even for a system with homogeneous servers. One known approximation idea in homogeneous service cases that works quite well for the design of services is the asymptotic equivalence in heavy traffic resource pooling cases [Harrison and López, 1999].

The purpose of this chapter is to get a deeper understanding about whether pooling strategies can be implemented to attain a higher service performance and a better collaboration income in real-life systems. In Chapters 2 and 3, we analyzed the cost-sharing problem in service pooling collaboration with general service times and impatient customers in the 'super-server' pooling setting. This assumption is also accepted in many other related service pooling studies [Anily and Haviv, 2010, Anily and Haviv, 2014, Yu et al., 2015]. Motivated by applications to real-life systems, it is reasonable to extend our previous research in multi-server systems [Özen et al., 2011, Karsten et al., 2015b], especially in queueing systems accounting for customer abandonment from queue ow-

ing to customer impatience. Such systems are widely used in service system design, e.g., the many-server heavy-traffic queueing model in call center applications [Dai and He, 2010, Jouini et al., 2013], the emergency hospital modeling problem [Gunal, 2012]. Therefore, we consider the stationary M/GI/S+M queuing model to describe the multi-server service system. In order to evaluate the 'super-server' assumptions, we use the M/GI/1+M modeling for the individual and the 'super-server' pooling system. Given the complexity of the analysis of the M/GI/S+M queue, we conduct a simulation study to compare between the two pooling situations.

The remainder of this chapter is structured as follows. We start in Section 4.2 by defining all notations of our models and deriving some comparisons for the pooling gains. Next, in Section 4.3, we provide a brief analysis of the relative service pooling game in case of the markovian service process. In Section 4.4, we numerically analyze the distributed costs with the impact of the variability of service times and the customer arrival rates for the two games. In Section 4.5, we consider the impact of the customer variability on service pooling gains. Finally, we give concluding comments.

## 4.2. Models and preliminary study

### 4.2.1. Service pooling modeling

We consider a set of $N = \{1, ..., n\}$, $n \in \mathbb{N}^+$ independent single-server service providers, a provider is denoted by $i \in N$. We address here the individual setting (without pooling), the 'super-server' setting (a single-server with a high service rate) and the multi-server setting (a pooled queue with parallel servers).

#### Individual service system $i$

In the individual setting, each provider is associated with its own Poisson customer arrival stream of rate $\lambda_i$, which does not affect other providers. All customers in the set $N$ are assumed to be classified in a same class, i.e., same priority for service. The waiting

space is assumed to be large enough such that no customer would be refused by the server. We assume that all the servers are identical, and service times are assigned to servers, which are i.i.d. and follow a general distribution with mean $1/\mu$ and coefficient of variation $cv$. The first come, first served (FCFS) discipline is used. Let customers be impatient. Patience times are assumed to be i.i.d. and exponentially distributed with mean $1/\theta$. If a customer has waited in the queue longer than his patience limit, he quits the waiting queue without a service. We assume that $\theta \leq \mu$ to limit the abandonment rate. Following these arguments, each service provider can be modeled as an M/GI/1+M queueing system (Figure 4.1), denoted by $Sys_i$, $i \in N$.

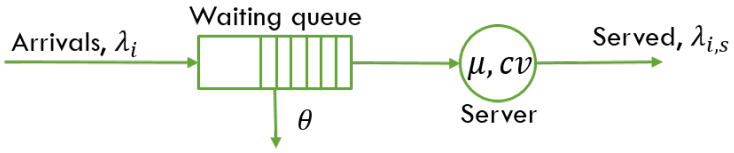



**Figure 4.1:** *An individual service provider i, an M/GI/1+M queue*

### *'Super-server'* pooling coalition $U$

For a subset $U = \{1, ..., u\} \subseteq N$ of $u$ service providers, we assume that they could operate their service capacities together to form a 'super-server'. The pooling system has a Poisson customer arrival stream of rate $\lambda_U = \sum_{i \in U} \lambda_i$. The waiting space is large enough, and FCFS is employed. In the pooling set, we assume that the overall service capacity is equivalent to the sum of the capacities of the $u$ individual providers in $U$. Consequently, service times are assigned to the pooled server, i.i.d. and generally distributed with mean $1/\mu_U$, $\mu_U = u\mu$ and coefficient of variation $cv$. Patience times are assigned to customers, i.i.d. and exponentially distributed with mean $1/\theta$. The $u$ providers are combined into a new M/GI/1+M queueing system (Figure 4.2), denoted by $Sys_U^s$, $U \subseteq N$. We have used a similar definition to define resource pooling in Chapters 2 and 3 for any given $\mu_i$ for server $i$.

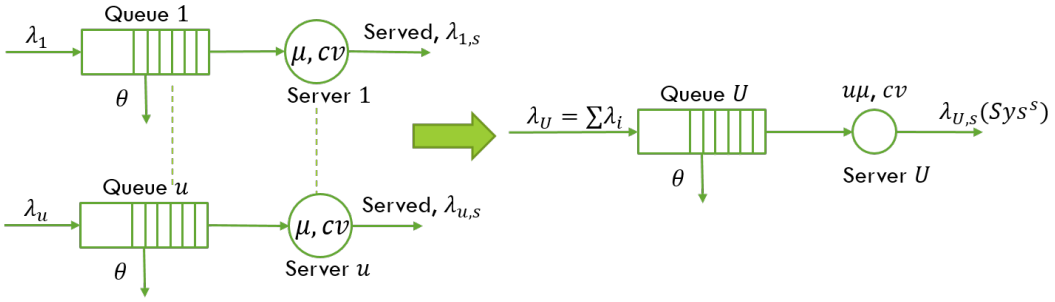



**Figure 4.2:** *'super-server' pooling concept in $U = \{1, ...u\}$, an M/GI/1+M queue*

### Multi-server pooling coalition $U$

We consider a subset $U = \{1, ..., u\} \subseteq N$ of $u$ service providers, they could operate their servers together to provide same services for all customers. We assume that all the $u$ servers are parallel and identical to all customers in $U$. Using similar assumptions in 'super-server' setting, the $u$ providers are combined into an M/GI/S+M queueing system. The customer arrival stream follows a Poisson process with rate $\lambda_U = \sum_{i \in U} \lambda_i$, and patience times are i.i.d. and exponentially distributed with mean $1/\theta$. Service times are assigned to servers, i.i.d. and general distributed with mean $1/\mu$ and coefficient of variation $cv$. The multi-server pooling coalition $U$ is shown in Figure 4.3, denoted by $Sys^m_U$, $U \subseteq N$.

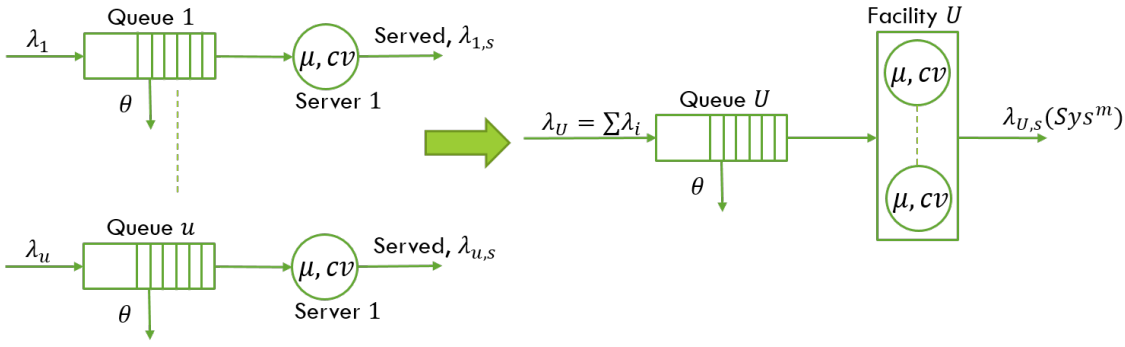



**Figure 4.3:** *Multi-server pooling concept in $U = \{1, ...u\}$, an M/GI/S+M queue*

**Cost functions**

We first denote by $L_q(i)$, $L_q^s(U)$, $L_q^m(U)$ and $P_a(i)$, $P_a^s(U)$, $P_a^m(U)$ the mean queue lengths and the mean customer abandonment probabilities for the three above setting, respectively. The total system cost of a service provider $i$ in the individual setting should consist of the server cost (resources cost measured by service capacity $C_h(i) = c_h\mu$), the queue cost (cost of customers waiting in the queue $C_w = c_w L_q(i)$) and the customer lost cost (loss for abandoned customers $C_a = c_a\lambda_i P_a(i)$). We give below the cost of an individual provider $i$,

$$C(i) = C_h(i) + C_w(i) + C_a(i) = c_h\mu + c_w L_q(i) + c_a\lambda_i P_a(i). \tag{4.1}$$

With similar definitions in the settings of $Sys^s$ and $Sys^m$, we define two relative costs of a subset $U$,

$$\begin{aligned} C^s(U) &= C_h^s(U) + C_w^s(U) + C_a^s(U) = c_h u\mu + c_w L_q^s(U) + c_a\lambda_U P_a^s(U), \\ C^m(U) &= C_h^m(U) + C_w^m(U) + C_a^m(U) = c_h u\mu + c_w L_q^m(U) + c_a\lambda_U P_a^m(U). \end{aligned} \tag{4.2}$$

Before discussing the results of pooling games, we first investigate in the next section, the relationship between the two pooling situations: $Sys^s$ and $Sys^m$.

### 4.2.2. Performance comparison

For simplification, we assume in this section that service times are exponentially distributed. Therefore, we compare the M/M/1+M queuing model $(\lambda, u\mu, \theta)$ with the M/M/S+M queueing model $(\lambda, \mu, \theta, u)$. Markov chains of the two compared models are presented in Figures 4.4 and 4.5.

**Lemma 4.1.** The expected number of customers in a pooled system with the setting $Sys^s$ is smaller than that in $Sys^m$.

*Proof.* Consider an M/M/1+M queuing model $(\lambda, u\mu, \theta)$ and an M/M/S+M queueing

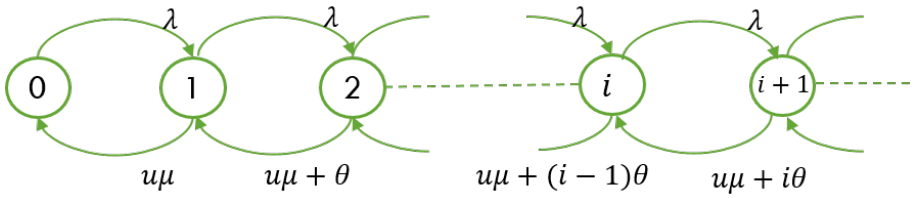



**Figure 4.4:** *The Markov process, an M/M/1+M queueing model*

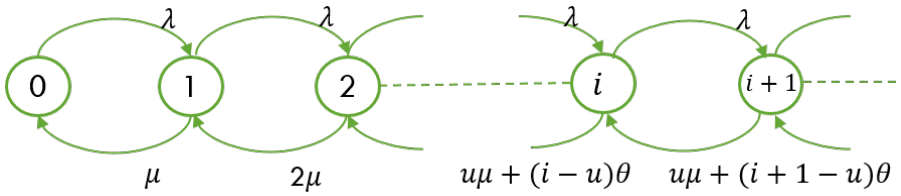



**Figure 4.5:** *The Markov process, an M/M/S+M queueing model*

model $(\lambda, \mu, \theta, u)$. We denote the number of customers in the two systems by the random variable $Y_{(\cdot)}$. The problem is to prove that

$$L_s^s(U) = \mathbb{E}(Y_s) \leq L_s^m(U) = \mathbb{E}(Y_m),$$

which is equivalent to

$$\frac{P(Y_s = k+1)}{Y_s = k} \leq \frac{P(Y_m = k+1)}{Y_m = k}, \text{ for } k \in \mathbb{N},$$

see page 208 in [Ferguson, 2014]. From Markov chains in Figures 4.4 and 4.5, we have

$$\begin{aligned}
\frac{P(Y_s = k+1)}{Y_s = k} &= \frac{\lambda}{u\mu + k\theta} \leq \frac{\lambda}{k\mu} = \frac{P(Y_m = k+1)}{Y_m = k}, \text{ for } 0 \leq k < \mu; \\
\frac{P(Y_s = k+1)}{Y_s = k} &= \frac{\lambda}{u\mu + k\theta} \leq \frac{\lambda}{u\mu + (k+1-u)\theta} = \frac{P(Y_m = k+1)}{Y_m = k}, \text{ for } k \geq \mu.
\end{aligned}$$

Thus, the expected number of customers $L_s$ in the M/M/1+M system is smaller, which completes the proof of the proposition. □

From Lemma 4.1, one can see that the overall expected number of customers in the setting $Sys^s$ is smaller than that in $Sys^m$, without any limit for the system parameters

$\{\lambda, \mu, \theta, u)\}$. The comparison of the queue length could be illustrated with a similar reasoning. We denote the number of customers in the two systems by the random variable $Y^-_{(\cdot)}$. Thus, we have identical $P(Y^-_{(\cdot)} = k+1)/P(Y^-_{(\cdot)} = k)$ for all $k \in \mathbb{N}^+$. The only difference is the value at $k = 0$. With a simple calculation, we can obtain Lemma 4.2

**Lemma 4.2.** The queue length in the setting $Sys^s$ is larger than that in $Sys^m$.

From the customer's point of view, we see that customers spend less overall time in $Sys^s$ but wait more time in queue. This phenomenon is well known in queueing systems without customer abandonment [Monahan, 2000]. When customers are impatient, $Sys^s$ loses more customers than $Sys^m$ dose. This reduces the difference between $L^s_q$ and $L^m_q$.

## 4.3. Service pooling games

In order to analyze the pooling strategies, we construct two corresponding service pooling games. We assume that every service provider $i$ (player $i$) in the set $N$ (the grand coalition) is economically independent. Suppose that a subset $U$, $\emptyset \subseteq U \subseteq N$, forms a coalition. In the coalition, providers operate their service resources together with pooling strategies. In above section, we assume that the service capacity in the pooling system is the sum of the individuals'. Thus, the server cost $C^{(\cdot)}_h$ is additive, which is meaningless in cooperative game study. The service pooling gain is only captured by $C^{(\cdot)}_w$ and $C^{(\cdot)}_a$. Consider that the abandonment process is Markovian. Therefore, for any coalition $U \subseteq N$, we denote the cost functions by

$$\begin{aligned} &C^s(\emptyset) = 0,\ C^s(U) = C^s_w(U) + C^s_a(U) = (c_w + c_a\theta)L^s_q(U) \in \mathbb{R}^+; \\ &C^m(\emptyset) = 0,\ C^m(U) = C^m_w(U) + C^m_a(U) = (c_w + c_a\theta)L^m_q(U) \in \mathbb{R}^+, \end{aligned} \tag{4.3}$$

under corresponding service pooling strategy, respectively. The pairs $(N, C^s)$ and $(N, C^m)$ define two cooperative games with transferable utility (TU-game). In particular, $C^s(\cdot)$ and $C^m(\cdot)$ are theirs characteristic functions (c.f.).

### 4.3.1. Service pooling games with exponential services

Now, we assume that service times are i.i.d. and exponentially distributed, and denote the relative games by $(N, C_M^{(\cdot)})$. Consider the Markov chains as shown in Figures 4.4 and 4.5. The expected queue length of the two settings could be calculated as follow.

$$L_q^s = \sum_{j=0}^{\infty}(j-1)\frac{a_j}{\sum_{l=0}^{\infty} a_l},\ a_0 = 1,\ a_j = \frac{\lambda^j}{\prod_{l=0}^{j-1}(u\mu + l\theta)},\ \text{for } j \in \mathbb{N}^+; \tag{4.4}$$

$$L_q^m = \sum_{j=0}^{\infty}(j-1)\frac{b_j}{\sum_{l=0}^{\infty} a_l},\ b_0 = 1,\ b_j = \frac{\lambda^j}{j!\mu^j},\ \text{if } j \le u; \tag{4.5}$$

$$b_j = \frac{\lambda^j}{u!\mu^u \prod_{l=u+1}^{j-1}(u\mu + (l-u)\theta)},\ \text{if } j > u.$$

In Chapters 2 and 3, we has shown that the social gains of $(N, C_M^s)$ is maximized when the grand coalition is formed, and the core is not empty. From Equation (4.4), it is easy to observe that the c.f. of $(N, C_M^s)$ is continuous in all parameters, $\{\lambda, u\mu, \theta\}$. However, as shown in Equation (4.5), it is not true for the multi-server setting, $\{\lambda, \mu, u, \theta\}$, i.e., $u \in \mathbb{N}^+$. This makes the theoretical analysis complicated. Next, we provide an example to illustrate the similarity of the two games.

### 4.3.2. Comparison of costs in $(N, C^s)$ and $(N, C^m)$

**Example.** Suppose that there are three players $N = \{1, 2, 3\}$, with $\mu = 10$, $c_w = c_a = 1$, and $\lambda_1 = 9, \lambda_2 = 5, \lambda_3 = 2$. As shown in Chapter 3, this example of the 'super-server' game $(N, C_M^s)$ is not concave and the Shapley value, denoted by $sh$, is not situated in the core. When $\theta$ increases, $sh$ becomes stable in an interval and moves out of the core again with a high abandonment. Reconsider it in the corresponding multi-server game $(N, C_M^m)$. The associated coalition costs and Shapley values of games $(N, C_M^s)$ and $(N, C_M^m)$ are presented in Table 4.1.

When customers are patient ($\theta = 0$), the two games are not stable with the Shapley value and have the same two unsatisfied coalitions $\{1, 2\}$ and $\{1, 3\}$, with $e_{sh}(U) < 0$ (by

| Coalitions | $\theta = 0$ | | | | $\theta = 0.5$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $C_M^s(U)$ | $\sum_{i\in U} sh_i^s$ | $C_M^m(U)$ | $\sum_{i\in U} sh_i^m$ | $C_M^s(U)$ | $\sum_{i\in U} sh_i^s$ | $C_M^m(U)$ | $\sum_{i\in U} sh_i^m$ |
| $\{1,2,3\}$ | 0.6095 | 0.6095 | 0.3130 | 0.3130 | 0.8130 | 0.8130 | 0.4135 | 0.4135 |
| $\{1,2\}$ | 1.6333 | 2.2240 | 1.3451 | 1.9779 | 1.7520 | 1.5285 | 1.4185 | 1.2139 |
| $\{2,3\}$ | 0.1885 | -2.5234 | 0.0977 | -2.6700 | 0.2597 | -0.7218 | 0.1342 | -0.9326 |
| $\{1,3\}$ | 0.6722 | 1.5184 | 0.4771 | 1.3189 | 0.8447 | 0.5958 | 0.5940 | 0.5457 |

**Table 4.1:** *Coalition and distributed costs with* $\theta = \{0, 0.5\}$ *for* $(N, C_M^s)$ *and* $(N, C_M^m)$

comparing data of Table 4.1). By increasing $\theta$, we numerically observe that the Shapley value is situated in the core for the two games. This phenomenon appears also with another services and the cases with higher number of players. In the next section, we give a systemic numerical comparison for three players games.

## 4.4. Numerical examples

We perform a simulation study in order to evaluate and compare the service pooling cost-sharing problem of the two strategies. The main reason of the simulation lies in the analytical complexity of the M/GI/S+M analysis and the intractable theoretical treatment of the related pooling games. Simulation is done with Matlab. We simulate the two pooling settings in three players set with deterministic, Erlang-2 ($cv = 1/\sqrt{2} = 0.707$) and hyperexponential (i.e., $\mu_{h,1} = \{0.5\mu + 1.5\mu\}$, $p_1 = \{0.25, 0.75\}$, $cv = 1.291$ and $\mu_{h,2} = \{0.2899\mu + 2.5798\mu\}$, $p_2 = \{0.2, 0.8\}$, $cv = 2$) services and calculate the corresponding exponential situation. From the identical servers assumption, i.e., same $\mu$, the individual utilization $\rho_i = \lambda_i/\mu$ of each server depends on own customer incoming rate. Therefore, we initially consider the case $\mu = 10$ and $\lambda = \{1, ...9\}$. We choose 20 cases of triples $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$ to evaluate the cost allocations using the Shapley value $sh$ and the nucleolus $\varphi^n$. The overall cost of each individual player or coalition is calculated by Equation (4.3) with $c_w = c_a = 1$. We choose this setting because the contrast between different customer incoming rates is the biggest reason for cost distribution and three

players game is a typical case for cost-sharing analysis. For evaluating the impact of customer abandonment, we choose four values $\theta = \{0, 0.5, 1, 3\}$. This gives us 80 different configurations for each game for each service distribution.

In all numerical tests, we find that the nucleolus is always situated in the core. It presents the non-emptiness of the core for all numerical examples. It means that the grand coalition is the most profitable coalition structure, and the three service providers would agree on pooling their service resources. Given the relationship shown for Markovian systems, the overall costs of $Sys^s$ are always larger than $Sys^m$ in all cases. Considering this difference between the two pooling settings, we calculate the individual saving in overall pooling saving by percentage in the grand coalition $N = \{1, 2, 3\}$, $sp_{x,i}^{(\cdot)} = [C^{(\cdot)}(\{i\}) - x_i^{(\cdot)}]/[\sum_{i \in N} C^{(\cdot)}(\{i\}) - C^{(\cdot)}(N)]\%$, i.e., $\mathbf{x} = (x_1, ..., x_n)$ present a cost allocation in $N$. Therefore, we give some cases in Tables 4.2. The first three cases are the cases with heterogeneous $\boldsymbol{\lambda}$, and homogeneous $\boldsymbol{\lambda}$ are selected in last three cases. 'Y'(Yes), 'N'(No),'N-Y'(No-Yes) indicate the stability of $\boldsymbol{\varphi}$, and 'N-Y' means that the stability changes with service times variability in these cases.

### 4.4.1. Impact of customer arrival rates $\lambda_i$

The $\lambda_i$ is the only individual parameter for each service provider $i$. Thus, its heterogeneity leads to heterogeneity in allocated costs. From Table 4.2, it is shown that the saving percentages using both the Shapley value or the Nucleolus allocations of $Sys^m$ and $Sys^s$ pooling systems are very closed in all the cases. The stabilities of the *sh* for the two pooling concepts are also very similar in the majority of cases, which varies with the service variability *cv* and the customer abandonment rate $\theta$. The special situations appear for the cases of $(1, 1, 2)$, in which all the three players have low initial utilizations. It is also shown that the *sh* is not stable for these cases with heterogeneous $\lambda_i$ when $\theta = 0$, i.e., $(1, 2, 9)$ with two small and one large $\lambda_i$, and $(1, 5, 9)$ with three heterogeneous $\lambda_i$. However, the *sh* is stable in the case $(1, 8, 9)$, i.e., one small and two large $\lambda_i$.

| $\lambda_i \in N$ | $cv$ | $\theta = 0$ | | | | $\theta = 0.5$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $sp^s_{sh}$% | $sp^m_{sh}$% | $sp^s_n$% | $sp^m_n$% | $sp^s_{sh}$% | $sp^m_{sh}$% | $sp^s_n$% | $sp^m_n$% |
| $(1,2,9)$ | 0 | 17.93% | 17.88% | 2.71% | 2.54% | 19.85% | 19.37% | 6.80% | 6.05% |
| | | 17.08% | 17.11% | 1.86% | 1.77% | 17.67% | 18.14% | 4.62% | 4.82% |
| | | 64.99% | 65.00% | 95.44% | 95.68% | 62.48% | 62.49% | 88.58% | 89.14% |
| | 0.707 | 17.84% | 17.81% | 2.51% | 2.39% | 20.27% | 19.79% | 7.71% | 6.89% |
| | | 17.04% | 17.11% | 1.70% | 1.69% | 17.88% | 18.40% | 5.33% | 5.50% |
| | | 65.12% | 65.09% | 95.79% | 95.91% | 61.86% | 61.82% | 86.96% | 87.60% |
| | 1 | 17.90% | 17.81% | 2.64% | 2.44% | 20.71% | 20.32% | 8.74% | 7.90% |
| | | 17.06% | 17.16% | 1.79% | 1.79% | 18.13% | 18.47% | 6.16% | 6.04% |
| | | 65.04% | 65.03% | 95.57% | 95.76% | 61.16% | 61.21% | 85.10% | 86.06% |
| | 1.291 | 17.94% | 17.93% | 2.72% | 2.64% | 21.44% | 20.78% | 10.23% | 8.88% |
| | | 17.07% | 17.14% | 1.85% | 1.85% | 18.25% | 18.70% | 7.03% | 6.79% |
| | | 65.00% | 64.93% | 95.44% | 95.52% | 60.30% | 60.52% | 82.74% | 84.33% |
| | 2 | 18.19% | 17.71% | 3.12% | 2.33% | 22.57% | 21.08% | 12.62% | 9.69% |
| | | 16.98% | 17.28% | 1.91% | 1.89% | 18.57% | 19.32% | 8.62% | 7.93% |
| | | 64.83% | 65.02% | 94.97% | 95.78% | 58.86% | 59.60% | 78.75% | 82.38% |
| *Stability* | | N | N | Y | Y | N | N | Y | Y |
| $(1,5,9)$ | 0 | 21.25% | 21.04% | 8.09% | 7.92% | 26.19% | 25.41% | 17.07% | 16.23% |
| | | 17.16% | 17.61% | 3.14% | 3.53% | 18.45% | 19.64% | 7.49% | 8.45% |
| | | 61.59% | 61.35% | 88.77% | 88.55% | 55.36% | 54.95% | 75.43% | 75.31% |
| | 0.707 | 20.92% | 20.90% | 7.53% | 7.67% | 27.53% | 26.45% | 19.48% | 18.24% |
| | | 17.18% | 17.57% | 2.97% | 3.40% | 18.66% | 20.12% | 8.46% | 9.57% |
| | | 61.90% | 61.53% | 89.50% | 88.92% | 53.81% | 53.43% | 72.05% | 72.19% |
| | 1 | 21.13% | 20.83% | 7.87% | 7.60% | 28.05% | 26.87% | 20.55% | 19.13% |
| | | 17.16% | 17.71% | 3.08% | 3.56% | 19.05% | 20.53% | 9.20% | 10.30% |
| | | 61.72% | 61.45% | 89.05% | 88.84% | 52.91% | 52.59% | 70.25% | 70.57% |
| | 1.291 | 21.56% | 21.22% | 8.61% | 8.29% | 29.04% | 28.01% | 22.30% | 21.14% |
| | | 17.11% | 17.71% | 3.24% | 3.75% | 19.26% | 20.74% | 10.11% | 11.26% |
| | | 61.33% | 61.06% | 88.15% | 87.96% | 51.69% | 51.24% | 67.60% | 67.60% |
| | 2 | 21.15% | 20.73% | 7.86% | 7.44% | 29.21% | 28.01% | 23.21% | 21.85% |
| | | 17.10% | 17.79% | 3.07% | 3.61% | 20.69% | 22.49% | 11.81% | 13.33% |
| | | 61.75% | 61.48% | 89.07% | 88.95% | 50.10% | 49.49% | 64.98% | 64.82% |
| *Stability* | | N | N | Y | Y | N-Y | N-Y | Y | Y |
| $(1,8,9)$ | 0 | 29.50% | 28.94% | 25.67% | 24.54% | 34.58% | 33.64% | 35.82% | 33.94% |
| | | 23.85% | 24.40% | 14.37% | 15.47% | 27.07% | 27.85% | 20.81% | 22.37% |
| | | 46.65% | 46.66% | 59.96% | 59.98% | 38.35% | 38.51% | 43.37% | 43.69% |
| | 0.707 | 29.91% | 28.92% | 26.42% | 24.51% | 35.38% | 33.68% | 37.43% | 34.02% |
| | | 23.02% | 23.85% | 12.85% | 14.36% | 27.11% | 28.38% | 20.89% | 23.42% |
| | | 47.07% | 47.23% | 60.74% | 61.12% | 37.51% | 37.95% | 41.68% | 42.56% |
| | 1 | 29.31% | 28.71% | 25.29% | 24.10% | 35.55% | 34.07% | 37.76% | 34.82% |
| | | 23.89% | 24.52% | 14.46% | 15.71% | 27.45% | 28.51% | 21.56% | 23.69% |
| | | 46.79% | 46.77% | 60.25% | 60.20% | 37.00% | 37.41% | 40.67% | 41.49% |
| | 1.291 | 30.20% | 28.89% | 27.07% | 24.46% | 35.43% | 35.15% | 37.53% | 36.96% |
| | | 23.38% | 24.48% | 13.42% | 15.63% | 28.55% | 28.86% | 23.77% | 24.38% |
| | | 46.42% | 46.62% | 59.50% | 59.91% | 36.02% | 36.00% | 38.70% | 38.66% |
| | 2 | 29.74% | 27.49% | 26.14% | 21.65% | 36.26% | 33.57% | 39.20% | 33.81% |
| | | 24.76% | 26.24% | 16.19% | 19.15% | 28.42% | 29.96% | 23.50% | 26.59% |
| | | 45.50% | 46.27% | 57.67% | 59.20% | 35.32% | 36.47% | 37.30% | 39.60% |
| *Stability* | | Y | Y | Y | Y | Y | Y | Y | Y |

**Table 4.2:** (a). *Heterogeneous $\lambda_i$ three players games*

| $\lambda_i \in N$ | $cv$ | $\theta = 1$ | | | | $\theta = 3$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $sp^s_{sh}\%$ | $sp^m_{sh}\%$ | $sp^s_n\%$ | $sp^m_n\%$ | $sp^s_{sh}\%$ | $sp^m_{sh}\%$ | $sp^s_n\%$ | $sp^m_n\%$ |
| $(1,2,9)$ | 0 | 20.57% | 20.36% | 8.61% | 8.01% | 22.35% | 21.64% | 12.41% | 10.95% |
| | | 18.16% | 18.42% | 6.20% | 6.07% | 18.95% | 19.51% | 9.01% | 8.83% |
| | | 61.26% | 61.22% | 85.18% | 85.92% | 58.70% | 58.84% | 78.58% | 80.22% |
| | 0.707 | 21.51% | 20.79% | 10.44% | 8.99% | 22.89% | 22.24% | 13.70% | 12.21% |
| | | 18.44% | 18.87% | 7.36% | 7.06% | 19.49% | 20.05% | 10.30% | 10.01% |
| | | 60.05% | 60.34% | 82.20% | 83.95% | 57.62% | 57.71% | 76.00% | 77.77% |
| | 1 | 21.72% | 21.21% | 10.95% | 9.85% | 23.42% | 22.68% | 14.72% | 13.15% |
| | | 18.62% | 19.06% | 7.85% | 7.71% | 19.52% | 20.26% | 10.82% | 10.73% |
| | | 59.67% | 59.73% | 81.20% | 82.44% | 57.06% | 57.06% | 74.46% | 76.12% |
| | 1.291 | 22.22% | 21.35% | 12.14% | 10.30% | 23.85% | 23.10% | 15.36% | 14.04% |
| | | 18.77% | 19.38% | 8.69% | 8.33% | 19.54% | 20.53% | 11.05% | 11.48% |
| | | 59.01% | 59.27% | 79.17% | 81.37% | 56.60% | 56.36% | 73.59% | 74.48% |
| | 2 | 22.59% | 22.06% | 13.34% | 11.85% | 24.02% | 23.18% | 16.44% | 14.64% |
| | | 19.71% | 20.09% | 10.47% | 9.88% | 20.45% | 21.46% | 12.87% | 12.92% |
| | | 57.71% | 57.85% | 76.19% | 78.28% | 55.53% | 55.36% | 70.69% | 72.44% |
| *Stability* | | N-Y | N | Y | Y | N-Y | N-Y | Y | Y |
| $(1,5,9)$ | 0 | 27.90% | 27.19% | 20.32% | 19.54% | 30.32% | 28.84% | 25.08% | 23.17% |
| | | 19.16% | 20.36% | 9.30% | 10.39% | 20.60% | 22.25% | 12.29% | 13.53% |
| | | 52.93% | 52.46% | 70.39% | 70.07% | 49.08% | 48.91% | 62.63% | 63.30% |
| | 0.707 | 28.99% | 27.87% | 22.48% | 21.16% | 31.32% | 29.08% | 27.04% | 24.01% |
| | | 19.92% | 21.27% | 10.83% | 11.71% | 21.24% | 23.24% | 13.69% | 14.81% |
| | | 51.09% | 50.85% | 66.69% | 67.12% | 47.43% | 47.67% | 59.27% | 61.18% |
| | 1 | 29.58% | 28.08% | 23.54% | 21.67% | 31.64% | 29.49% | 27.66% | 24.85% |
| | | 19.91% | 21.69% | 11.09% | 12.37% | 21.32% | 23.67% | 13.90% | 15.60% |
| | | 50.50% | 50.23% | 65.38% | 65.96% | 47.03% | 46.84% | 58.44% | 59.56% |
| | 1.291 | 30.29% | 28.01% | 24.88% | 21.83% | 32.66% | 29.52% | 29.49% | 25.15% |
| | | 20.27% | 22.30% | 11.93% | 12.96% | 21.15% | 24.27% | 13.97% | 16.34% |
| | | 49.44% | 49.70% | 63.19% | 65.21% | 46.19% | 46.20% | 56.54% | 58.51% |
| | 2 | 31.28% | 29.18% | 26.99% | 24.20% | 32.17% | 29.69% | 28.99% | 25.77% |
| | | 20.83% | 23.45% | 12.79% | 15.20% | 22.29% | 25.30% | 15.29% | 17.82% |
| | | 47.89% | 47.38% | 60.22% | 60.60% | 45.54% | 45.01% | 55.72% | 56.41% |
| *Stability* | | Y | Y | Y | Y | Y | Y | Y | Y |
| $(1,8,9)$ | 0 | 35.89% | 33.98% | 38.44% | 34.63% | 35.90% | 34.57% | 38.48% | 35.81% |
| | | 27.18% | 28.39% | 21.03% | 23.44% | 28.47% | 29.48% | 23.61% | 25.64% |
| | | 36.93% | 37.63% | 40.53% | 41.93% | 35.62% | 35.94% | 37.92% | 38.55% |
| | 0.707 | 36.01% | 34.53% | 38.69% | 35.72% | 36.69% | 34.68% | 40.04% | 36.02% |
| | | 28.02% | 28.90% | 22.70% | 24.46% | 28.51% | 29.80% | 23.68% | 26.27% |
| | | 35.97% | 36.58% | 38.61% | 39.82% | 34.80% | 35.52% | 36.27% | 37.71% |
| | 1 | 36.16% | 34.39% | 38.99% | 35.44% | 36.88% | 34.43% | 40.42% | 35.52% |
| | | 27.94% | 29.16% | 22.55% | 25.00% | 28.51% | 30.11% | 23.69% | 26.88% |
| | | 35.90% | 36.45% | 38.47% | 39.56% | 34.61% | 35.46% | 35.89% | 37.60% |
| | 1.291 | 35.99% | 34.65% | 38.65% | 35.96% | 37.27% | 34.70% | 41.21% | 36.06% |
| | | 27.84% | 29.06% | 22.34% | 24.79% | 28.61% | 30.31% | 23.89% | 27.28% |
| | | 36.17% | 36.29% | 39.01% | 39.25% | 34.12% | 34.99% | 34.90% | 36.65% |
| | 2 | 37.82% | 34.06% | 42.31% | 34.79% | 38.55% | 33.64% | 43.77% | 33.95% |
| | | 28.30% | 30.45% | 23.26% | 27.57% | 28.23% | 31.06% | 23.13% | 28.79% |
| | | 33.88% | 35.49% | 34.43% | 37.64% | 33.22% | 35.29% | 33.10% | 37.26% |
| *Stability* | | Y | Y | Y | Y | Y | Y | Y | Y |

**Table 4.2:** (b). *Heterogeneous $\lambda_i$ three players games*

| $\lambda_i \in N$ | $cv$ | $\theta = 0$ | | | | $\theta = 0.5$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $sp^s_{sh}$% | $sp^m_{sh}$% | $sp^s_n$% | $sp^m_n$% | $sp^s_{sh}$% | $sp^m_{sh}$% | $sp^s_n$% | $sp^m_n$% |
| $(1,1,2)$ | 0 | 25.94% | 25.73% | 18.60% | 18.11% | 25.45% | 25.61% | 17.56% | 17.97% |
| | | 25.94% | 25.73% | 18.60% | 18.11% | 25.45% | 25.61% | 17.56% | 17.97% |
| | | 48.20% | 48.49% | 62.90% | 63.75% | 49.12% | 48.72% | 65.19% | 64.06% |
| | 0.707 | 25.55% | 25.43% | 17.78% | 17.50% | 26.54% | 25.56% | 19.79% | 17.76% |
| | | 25.55% | 25.43% | 17.78% | 17.50% | 26.54% | 25.56% | 19.79% | 17.76% |
| | | 48.92% | 49.07% | 64.47% | 64.93% | 46.99% | 48.83% | 60.49% | 64.48% |
| | 1 | 25.74% | 25.38% | 18.20% | 17.35% | 26.07% | 25.53% | 18.71% | 17.72% |
| | | 25.74% | 25.38% | 18.20% | 17.35% | 26.07% | 25.53% | 18.71% | 17.72% |
| | | 48.54% | 49.33% | 63.62% | 64.93% | 48.06% | 48.88% | 62.80% | 64.56% |
| | 1.291 | 26.02% | 25.17% | 18.81% | 17.02% | 26.79% | 25.86% | 20.25% | 18.48% |
| | | 26.02% | 25.17% | 18.81% | 17.02% | 26.79% | 25.86% | 20.25% | 18.48% |
| | | 47.99% | 49.60% | 62.54% | 65.81% | 46.42% | 48.23% | 59.64% | 63.04% |
| | 2 | 26.08% | 25.45% | 18.91% | 17.52% | 27.22% | 25.77% | 21.13% | 18.19% |
| | | 26.08% | 25.45% | 18.91% | 17.52% | 27.22% | 25.77% | 21.13% | 18.19% |
| | | 47.75% | 49.15% | 62.23% | 64.90% | 45.56% | 48.46% | 57.73% | 63.56% |
| *Stability* | | Y | N | Y | Y | Y | N | Y | Y |
| $(4,5,5)$ | 0 | 28.94% | 28.47% | 24.54% | 23.58% | 29.73% | 29.99% | 26.11% | 26.65% |
| | | 35.52% | 35.77% | 37.72% | 38.20% | 35.14% | 35.01% | 36.94% | 36.68% |
| | | 35.52% | 35.77% | 37.72% | 38.20% | 35.14% | 35.01% | 36.94% | 36.68% |
| | 0.707 | 29.30% | 29.26% | 25.26% | 25.18% | 29.53% | 30.20% | 25.75% | 27.06% |
| | | 35.34% | 35.38% | 37.37% | 37.41% | 35.23% | 34.91% | 37.13% | 36.47% |
| | | 35.34% | 35.38% | 37.37% | 37.41% | 35.23% | 34.91% | 37.13% | 36.47% |
| | 1 | 29.39% | 29.40% | 25.45% | 25.46% | 30.17% | 30.12% | 27.00% | 26.92% |
| | | 35.30% | 35.30% | 37.28% | 37.26% | 34.92% | 34.94% | 36.49% | 36.53% |
| | | 35.30% | 35.30% | 37.28% | 37.26% | 34.92% | 34.94% | 36.49% | 36.53% |
| | 1.291 | 28.99% | 29.83% | 24.64% | 26.34% | 31.83% | 30.91% | 30.32% | 28.50% |
| | | 35.51% | 35.08% | 37.68% | 36.83% | 34.09% | 34.54% | 34.84% | 35.75% |
| | | 35.51% | 35.08% | 37.68% | 36.83% | 34.09% | 34.54% | 34.84% | 35.75% |
| | 2 | 29.97% | 30.24% | 26.62% | 27.14% | 30.69% | 30.00% | 28.04% | 26.66% |
| | | 35.01% | 34.88% | 36.69% | 36.43% | 34.66% | 35.00% | 35.98% | 36.67% |
| | | 35.01% | 34.88% | 36.69% | 36.43% | 34.66% | 35.00% | 35.98% | 36.67% |
| *Stability* | | Y | Y | Y | Y | Y | Y | Y | Y |
| $(8,9,9)$ | 0 | 28.19% | 28.26% | 23.05% | 23.18% | 32.03% | 32.10% | 30.74% | 30.87% |
| | | 28.19% | 28.26% | 23.05% | 23.18% | 32.03% | 32.10% | 30.74% | 30.87% |
| | | 43.62% | 43.49% | 53.91% | 53.65% | 35.93% | 35.80% | 38.52% | 38.26% |
| | 0.707 | 28.07% | 28.09% | 22.82% | 22.84% | 32.13% | 32.16% | 30.92% | 30.99% |
| | | 28.07% | 28.09% | 22.82% | 22.84% | 32.13% | 32.16% | 30.92% | 30.99% |
| | | 43.85% | 43.83% | 54.37% | 54.32% | 35.75% | 35.68% | 38.16% | 38.02% |
| | 1 | 28.04% | 28.30% | 22.74% | 23.27% | 32.16% | 32.32% | 30.98% | 31.31% |
| | | 28.04% | 28.30% | 22.74% | 23.27% | 32.16% | 32.32% | 30.98% | 31.31% |
| | | 43.92% | 43.40% | 54.52% | 53.46% | 35.68% | 35.36% | 38.04% | 37.39% |
| | 1.291 | 28.36% | 28.30% | 23.39% | 23.27% | 32.35% | 33.10% | 31.36% | 32.87% |
| | | 28.36% | 28.30% | 23.39% | 23.27% | 32.35% | 33.10% | 31.36% | 32.87% |
| | | 43.28% | 43.39% | 53.22% | 53.45% | 35.30% | 33.79% | 37.27% | 34.25% |
| | 2 | 28.83% | 28.48% | 24.32% | 23.63% | 32.57% | 32.84% | 31.82% | 32.34% |
| | | 28.83% | 28.48% | 24.32% | 23.63% | 32.57% | 32.84% | 31.82% | 32.34% |
| | | 42.34% | 43.03% | 51.35% | 52.73% | 34.85% | 34.32% | 36.37% | 35.32% |
| *Stability* | | Y | Y | Y | Y | Y | Y | Y | Y |

**Table 4.2:** (c). *Homogeneous $\lambda_i$ three players games*

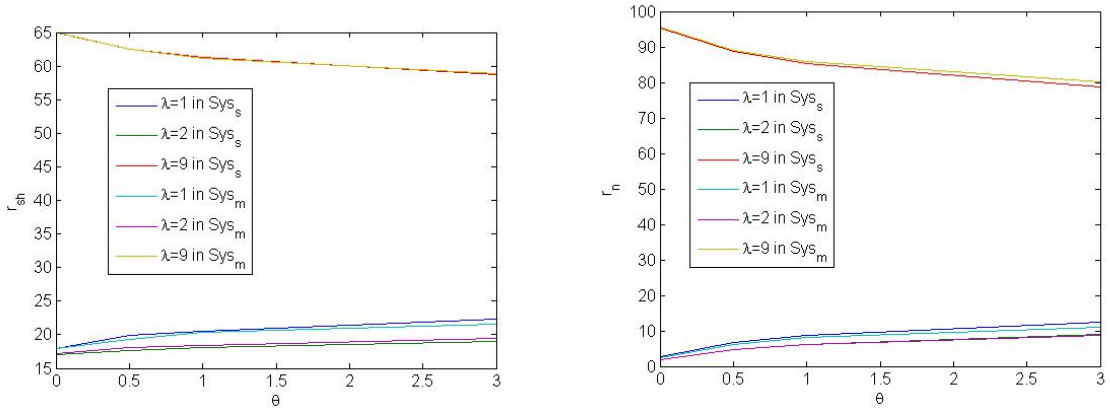
| $\lambda_i \in N$ | $cv$ | $\theta = 1$ | | | | $\theta = 3$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $sp^s_{sh}\%$ | $sp^m_{sh}\%$ | $sp^s_n\%$ | $sp^m_n\%$ | $sp^s_{sh}\%$ | $sp^m_{sh}\%$ | $sp^s_n\%$ | $sp^m_n\%$ |
| $(1,1,2)$ | 0 | 25.66% | 25.67% | 20.39% | 18.01% | 27.71% | 26.48% | 22.15% | 26.48% |
| | | 25.66% | 25.67% | 20.39% | 18.01% | 27.71% | 26.48% | 22.15% | 26.48% |
| | | 48.44% | 48.63% | 59.43% | 63.98% | 44.58% | 47.02% | 55.83% | 47.02% |
| | 0.707 | 26.28% | 26.04% | 19.12% | 18.73% | 26.44% | 25.78% | 19.51% | 25.78% |
| | | 26.28% | 26.04% | 19.12% | 18.73% | 26.44% | 25.78% | 19.51% | 25.78% |
| | | 47.41% | 47.94% | 61.58% | 62.54% | 47.12% | 48.44% | 60.88% | 48.52% |
| | 1 | 26.17% | 25.63% | 18.99% | 18.05% | 26.57% | 26.07% | 19.81% | 26.07% |
| | | 26.17% | 25.63% | 18.99% | 18.05% | 26.57% | 26.07% | 19.81% | 26.07% |
| | | 47.66% | 48.66% | 62.02% | 63.90% | 46.86% | 47.91% | 60.47% | 47.91% |
| | 1.291 | 27.17% | 25.61% | 21.03% | 17.93% | 27.47% | 26.83% | 21.64% | 26.83% |
| | | 27.17% | 25.61% | 21.03% | 17.93% | 27.47% | 26.83% | 21.64% | 26.83% |
| | | 45.62% | 48.80% | 57.89% | 64.14% | 45.00% | 46.37% | 56.73% | 46.37% |
| | 2 | 26.43% | 25.87% | 20.15% | 18.39% | 26.79% | 26.13% | 20.25% | 26.13% |
| | | 26.43% | 25.87% | 20.15% | 18.39% | 26.79% | 26.13% | 20.25% | 26.13% |
| | | 47.17% | 48.28% | 59.79% | 63.22% | 46.40% | 47.78% | 59.42% | 47.78% |
| *Stability* | | Y | N | Y | Y | Y | N-Y | Y | Y |
| $(4,5,5)$ | 0 | 30.37% | 30.06% | 27.41% | 26.79% | 30.70% | 30.73% | 28.06% | 28.12% |
| | | 34.80% | 34.97% | 36.29% | 36.60% | 34.65% | 34.63% | 35.98% | 35.94% |
| | | 34.80% | 34.97% | 36.29% | 36.60% | 34.65% | 34.63% | 35.98% | 35.94% |
| | 0.707 | 30.47% | 29.82% | 27.61% | 26.31% | 31.10% | 30.92% | 28.88% | 28.50% |
| | | 34.76% | 35.09% | 36.20% | 36.85% | 34.44% | 34.54% | 35.57% | 35.75% |
| | | 34.76% | 35.09% | 36.20% | 36.85% | 34.44% | 34.54% | 35.57% | 35.75% |
| | 1 | 30.53% | 30.47% | 27.72% | 27.60% | 31.11% | 31.02% | 28.89% | 28.71% |
| | | 34.74% | 34.77% | 36.14% | 36.20% | 34.45% | 34.49% | 35.55% | 35.65% |
| | | 34.74% | 34.77% | 36.14% | 36.20% | 34.45% | 34.49% | 35.55% | 35.65% |
| | 1.291 | 30.86% | 30.90% | 28.39% | 28.48% | 30.59% | 30.86% | 27.84% | 28.40% |
| | | 34.57% | 34.55% | 35.81% | 35.76% | 34.71% | 34.57% | 36.07% | 35.80% |
| | | 34.57% | 34.55% | 35.81% | 35.76% | 34.71% | 34.57% | 36.07% | 35.80% |
| | 2 | 30.79% | 30.62% | 28.25% | 27.91% | 32.17% | 31.54% | 31.01% | 29.74% |
| | | 34.61% | 34.69% | 35.88% | 36.04% | 33.92% | 34.23% | 34.50% | 35.13% |
| | | 34.61% | 34.69% | 35.88% | 36.04% | 33.92% | 34.23% | 34.50% | 35.13% |
| *Stability* | | Y | Y | Y | Y | Y | Y | Y | Y |
| $(8,9,9)$ | 0 | 32.14% | 32.10% | 30.95% | 30.87% | 32.41% | 32.76% | 31.49% | 32.20% |
| | | 32.14% | 32.10% | 30.95% | 30.87% | 32.41% | 32.76% | 31.49% | 32.20% |
| | | 35.71% | 35.79% | 38.09% | 38.26% | 35.18% | 34.47% | 37.02% | 35.61% |
| | 0.707 | 32.54% | 32.67% | 31.75% | 32.00% | 32.82% | 32.97% | 32.30% | 32.61% |
| | | 32.54% | 32.67% | 31.75% | 32.00% | 32.82% | 32.97% | 32.30% | 32.61% |
| | | 34.91% | 34.66% | 36.49% | 35.99% | 34.37% | 34.05% | 35.40% | 34.77% |
| | 1 | 32.45% | 32.60% | 31.56% | 31.87% | 32.71% | 32.86% | 32.09% | 32.39% |
| | | 32.45% | 32.60% | 31.56% | 31.87% | 32.71% | 32.86% | 32.09% | 32.39% |
| | | 35.10% | 34.80% | 36.87% | 36.26% | 34.57% | 34.28% | 35.82% | 35.22% |
| | 1.291 | 32.03% | 32.78% | 30.73% | 32.23% | 32.93% | 33.14% | 32.52% | 32.95% |
| | | 32.03% | 32.78% | 30.73% | 32.23% | 32.93% | 33.14% | 32.52% | 32.95% |
| | | 35.93% | 34.43% | 38.54% | 35.54% | 34.15% | 33.72% | 34.96% | 34.10% |
| | 2 | 33.62% | 33.23% | 33.90% | 33.13% | 33.41% | 32.64% | 33.49% | 31.94% |
| | | 33.62% | 33.23% | 33.90% | 33.13% | 33.41% | 32.64% | 33.49% | 31.94% |
| | | 32.77% | 33.53% | 32.20% | 33.73% | 33.18% | 34.72% | 33.03% | 36.12% |
| *Stability* | | Y | Y | Y | Y | Y | Y | Y | Y |

**Table 4.2:** (d). *Homogeneous $\lambda_i$ three players games*

Considering the unsatisfied coalitions for the example in Section 4.3.2, the service provider with a large $\lambda_i$ pays too much in the *sh*, which leads to the instability of the grand coalition. In $(1,8,9)$, there are two providers with relative large $\lambda_i$ sharing the overmuch cost. Although stability of the *sh* could not be ensured, the *sh* is more fair than the $\varphi^n$.

### 4.4.2. Impact of customer abandonment $\theta$

Apparently, the service systems with a high utilization are more sensitive to $\theta$ than those with a low utilization. In Table 4.2, it is shown that the difference between $sp^s_{(\cdot)}$ and $sp^m_{(\cdot)}$ does not vary with $\theta$. In the majority of cases (an example is shown in Figure 4.6) for the two pooling settings, the data clearly show that the saving percentage of the player with a small $\lambda_i$ is increasing with $\theta$ in $(0,3)$, in contrast to the player with a large $\lambda_i$. This is because the abandonment could reduce heterogeneity of the queue lengths among the service providers with the same service capacity, i.e., equal $u$ and $\mu$. The special situation $(1,1,2)$ of three service providers with low utlizations, is the case which is less sensitive to $\theta$.




(a) $sp_{sh}$ of the case $\boldsymbol{\lambda} = (1,2,9)$ in $\theta$ (b) $sp_n$ of the case $\boldsymbol{\lambda} = (1,2,9)$ in $\theta$

**Figure 4.6:** *Relative savings of the case* $\boldsymbol{\lambda} = (1,2,9)$ *in* $\theta$

### 4.4.3. Impact of service times variability *cv*

In all our cases, it is clear that all the costs for individual or pooling systems are increasing with *cv*. The allocated costs using the two allocations are also increasing with *cv*. From Table 4.2, it is shown that the saving percentage of the two allocations varies a little bit with *cv* in both two pooling settings. The stability of *sh* varies with *cv* under certain values of $\theta$. We give a more detailed information about 'N-Y' in Table 4.3.

| $\lambda_i \in N$ | $\theta$ | $Sys^{(\cdot)}$ | *cv* | | | | |
|---|---|---|---|---|---|---|---|
| | | | 0 | 0.707 | 1 | 1.291 | 2 |
| (1,2,9) | 1 | $Sys^s$ | N | N | N | N | Y |
| | 1 | $Sys^m$ | N | N | N | N | N |
| | 3 | $Sys^s$ | Y | Y | Y | Y | Y |
| | 3 | $Sys^m$ | N | Y | Y | Y | Y |
| (1,5,9) | 0.5 | $Sys^s$ | N | N | Y | Y | Y |
| | 0.5 | $Sys^m$ | N | N | N | Y | Y |

**Table 4.3:** *Stability of the Shapley value in cv*

From the *Pollaczek-Khinchine formula*, it is obvious that *cv* dose not impact the relative costs of the individual and coalition systems under the two pooling settings. When $\theta$ increases, all queue lengths increase in *cv* and the longer queues have more abandoned customers. Therefore, increasing *cv* could enhance the impact of $\theta$.

## 4.5. Pooling in multi-class service systems

We conduct a simulation study in order to analyze whether pooling heterogeneous service times allows service providers to obtain a better performance and the impact of customer patience. In this study, we consider two single-server service systems $\{1,2\}$ with different customer classes and construct three sets of experiment to illustrate the pooling gains. We assume that customer arrival process follows a Poisson process with rate $\lambda_i$, $i \in \{1,2\}$, respectively. Suppose that service times of one class are i.i.d. with

mean $\mu_i^{-1}$, $i \in \{1,2\}$ and not change in the pooling system. In the pooling setting, the two classes have the same priority and are served under FCFS discipline using two identical parallel servers. When customers are impatient, we assume that the two classes have an identical abandonment rate $\theta = 1$.

In each set, we consider three cases for the relationship between $\mu_1$ and $\mu_2$, i.e., $\mu_2 = p_{mc}\mu_1$, $p_{mc} = \{2,5,10\}$, which means that the first class generally gets a longer service time than the second one. We choose the relative reduction of queue length for each class and the overall service system to evaluate the pooling performance by using the following expression,

$$R_{Lq} = \begin{cases} \frac{L_{q,i} - L_{q,i}^{p}}{L_{q,i}}, \text{for class } i \in \{1,2\} \\ \frac{\sum L_{q,i} - L_q}{\sum L_{q,i}}, \text{for pooling } i \in \{1,2\} \end{cases} \tag{4.6}$$

Firstly, we consider two classes of customers with same server utilization and without customer abandonment. Our first set of experiments consists of twelve cases with four values of server utilisation. We consider $\rho_1 = \{0.2, 0.4, 0.6, 0.8\}$ with $\theta = 0$. The values of pooling gains, noted by $R_{Lq}$, in these twelve cases are given in Table 4.4. We find that customers of class 1 get more service quality improvement than those of class 2. When the class heterogeneity $p_{mc}$ increases, the relative pooling performance $R_{Lq}$ of class 2 and overall system $\{1,2\}$ decrease. On the contrary, $R_{Lq}$ of class 1 increases. This is because as $p_{mc}$ increases, service times of class 2 are much shorter than those of class 1. Thus, the congestion impact in the pooling system for each other is much important for class 2. For a same $p_{mc}$, $R_{Lq}$ decreases with utilizations $\rho_1 = \rho_2$. It is also the effect of congestion.

Secondly, we consider different initial utilisations $\rho_i = \lambda_i/\mu_i$ with $\rho_2 = p_{du}\rho_1$, $p_{du} = \{1/2, 3/2\}$ in Table 4.4. Thus, there are also twelve cases with different server utilisations considering two values of utilization in class 1, $\rho_1 = \{0.2, 0.6\}$. Considering the relative results in Table 4.4 , we could conclude that the relative gains of class 2 increases with customer arrivals augmentation in this class. Simultaneously, the relative gains of class 1 reduces with additional incoming customers of class 2. It is because that the heavy load

| $\rho_1$ | $p_{mc}=2$ | | | $p_{mc}=5$ | | | $p_{mc}=10$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | C-1 | C-2 | All | C-1 | C-2 | All | C-1 | C-2 | All |
| 0.2 | 88.00% | 77.40% | 82.70% | 91.20% | 61.60% | 76.40% | 94.20% | 37.40% | 65.80% |
| 0.4 | 78.10% | 57.44% | 67.77% | 86.46% | 32.20% | 59.33% | 84.40% | −39.09% | 22.66% |
| 0.6 | 75.06% | 46.92% | 60.99% | 80.00% | 4.86% | 42.43% | 80.46% | −92.24% | −5.89% |
| 0.8 | 70.74% | 41.11% | 55.93% | 73.93% | −28.98% | 22.47% | 76.40% | −134.66% | −29.13% |

**Table 4.4:** *Relative pooling performance of two classes of customers with same utilisation and* $\theta = 0$

impact on class 2 in individual setting is reduced with the joining of server $\{1\}$.

| $\theta=0$ | | $p_{mc}=2$ | | | $p_{mc}=5$ | | | $p_{mc}=10$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\rho_1$ | $p_{du}$ | C-1 | C-2 | All | C-1 | C-2 | All | C-1 | C-2 | All |
| 0.2 | 1/2 | 95.80% | 66.70% | 90.51% | 94.80% | 50.50% | 86.75% | 96.00% | −5.30% | 77.58% |
| - | 3/2 | 81.60% | 79.47% | 80.06% | 86.40% | 67.10% | 72.50% | 91.20% | 46.33% | 58.90% |
| 0.6 | 1/2 | 84.12% | −15.97% | 71.61% | 89.58% | −85.97% | 67.63% | 90.94% | −234.37% | 50.28% |
| - | 3/2 | 45.89% | 81.60% | 78.03% | 57.02% | 64.78% | 64.01% | 65.31% | 41.29% | 43.69% |

**Table 4.5:** *Relative pooling performance of two classes of customers with* $p_{du} = \{1/2, 3/2\}$ *and* $\theta = 0$

Lastly, we consider the impact of customer impatience in Tables 4.6 and 4.7. The set of experiments consists of twelve cases with abandonment rate $\theta = \{0.5, 1\}$ and identical settings for other parameters. It is shown that the impact of customer heterogeneity $p_{mc}$ still reduces the overall pooling gains and increases the difference of the relative gains between classes. However, comparing the relative performance under different $\theta$, we find that the impact of $\theta$ weaken the impact of $p_{mc}$. The negative individual $R_{L_q}$ of class 2, caused by class heterogeneity, is reduced with the presence of $\theta$.

Based on above discussions, we could conclude that the service pooling in multi-class systems is not always profitable owing to the high heterogeneity of classes and the high system utilization, which is also shown in real-life systems [Vanberkel et al., 2012].

| $\theta = 1$ | | $p_{mc} = 2$ | | | $p_{mc} = 5$ | | | $p_{mc} = 10$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\rho_1$ | $p_{du}$ | C-1 | C-2 | All | C-1 | C-2 | All | C-1 | C-2 | All |
| 0.2 | 1/2 | 89.44% | 81.54% | 87.35% | 90.00% | 38.96% | 74.71% | 92.78% | 23.40% | 68.98% |
| - | 3/2 | 70.56% | 71.85% | 71.54% | 82.22% | 68.28% | 70.63% | 85.56% | 57.03% | 61.45% |
| 0.6 | 1/2 | 70.49% | 26.22% | 57.78% | 69.79% | −11.59% | 38.46% | 71.62% | −101.02% | 1.04% |
| - | 3/2 | 28.59% | 41.01% | 38.29% | 39.58% | 35.72% | 36.21% | 50.85% | 26.51% | 28.69% |

**Table 4.6:** *Relative pooling performance of two classes of customers with* $p_{du} = \{1/2, 3/2\}$ *and* $\theta = \{1\}$

| $\theta = 0.5$ | | $p_{mc} = 2$ | | | $p_{mc} = 5$ | | | $p_{mc} = 10$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\rho_1$ | $p_{du}$ | C-1 | C-2 | All | C-1 | C-2 | All | C-1 | C-2 | All |
| 0.7 | 1/2 | 65.55% | 11.46% | 50.99% | 69.60% | −56.03% | 28.98% | 73.76% | −133.23% | 1.22% |
| - | 3/2 | 23.32% | 38.78% | 35.58% | 40.72% | 37.63% | 38.01% | 44.50% | 25.41% | 26.97% |
| 0.9 | 1/2 | 65.46% | −4.64% | 47.63% | 67.17% | −71.98% | 21.78% | 67.99% | −183.82% | −22.83% |
| - | 3/2 | 15.72% | 31.38% | 28.04% | 24.48% | 28.36% | 27.92% | 27.24% | 24.76% | 24.92% |
| $\theta = 1$ | | C-1 | C-2 | All | C-1 | C-2 | All | C-1 | C-2 | All |
| 0.7 | 1/2 | 62.41% | 10.00% | 47.12% | 66.48% | −32.26% | 29.42% | 70.96% | −97.27% | 0.24% |
| - | 3/2 | 31.03% | 41.07% | 38.84% | 40.59% | 33.72% | 34.61% | 46.49% | 26.23% | 27.94% |
| 0.9 | 1/2 | 57.99% | 3.38% | 41.45% | 60.76% | −44.51% | 18.81% | 63.41% | −124.20% | −21.62% |
| - | 3/2 | 18.98% | 30.97% | 28.34% | 25.34% | 25.87% | 25.80% | 33.01% | 24.75% | 25.32% |

**Table 4.7:** *Relative pooling performance of two classes of customers with high utilisations* $\rho_1 = \{0.7, 0.9\}$

## 4.6. Conclusion

We considered the service resource pooling problem while accounting for the important feature of customer abandonment, and investigated the cooperative strategy among independent service providers in both the 'super-server' pooling setting and the multi-server pooling setting. We assumed that all individual servers are identical and 'super-server' is $u$ (the number of servers in the coalition) times faster than the individual one. With markovian service distributions, we provided a brief analysis for the expected queue length (related to the abandonment probability) and the expected number of cus-

tomers in the system. Numerical experiments showed the similar impacts of the service variability and the customer abandonment on the cost allocations of the two games.

Under the multi-server assumptions, we numerically tested the resource pooling performance in terms of the expected queue length in a 2-class service system. With system congestion and heterogeneity of classes, the class with faster service is less qualified than before. When customers are impatient, this individual quality reduction is weakened.

# Chapter 5

# Conclusion and Perspectives

We formulated resource pooling games for three different service models and derived structure properties under the point of view of cooperation. Using cooperative game theory, we studied cooperative strategies among a set of independent service providers. Our approach consists of addressing two consecutive questions: 1) which coalition strategy should be used? and 2) which allocation rule should be selected in order to maintain the stability of the coalition?

In Chapter 2, we investigated the service pooling game for M/GI/1 service systems. When service capacities are fixed, we proved that the stable cost allocations always exist for the grand coalition. When service capacities are optimized to minimize the total operating costs, we analyzed the properties related to the optimal service rate. We presented a combined allocation policy in this situation.

In Chapter 3, we extended the pooling problem for service systems while accounting for the important feature of customer abandonment. When the service providers directly combine their queues and service capacities, we proved the existence of stable allocations. We studied the convexity of the expected queue length in the abandonment rate and found the special condition of the stability of the Shapley value.

In Chapter 4, we considered the multi-server pooling setting. We investigated numerically the impact of service duration variability and customer abandonment on the pooling game. We compared between cost-sharing results of the two resource pool-

ing concepts, with or without the 'super-server' assumptions. Finally, we analyzed the pooling of two-class service systems.

The results obtained in this Ph.D. thesis provide avenues for futures research. We next highlight some of them.

Concerning the work of Chapter 1, it would be interesting to extend the analysis to the cases where the parameters $c_h$, $c_w$ or/and $f$ depends on the service provider. For instance, consider the case of several hospital departments with a shared ward of beds. The cost of providing medical services by these departments may vary in a great deal, depending on the medical equipment as well as the medication. Furthermore, the cost of waiting for a bed may also be significantly different for patients that visit different departments. Our difficulty with different $c_{h,i}$ is how to define a reasonable pooling $c_{h,S}$, especially for the case of optimized service capacity. It is not appropriate to just use the average of the individual costs $c_{h,S} = \sum_S c_{h,i}\mu_i / \sum_S \mu_i$. We could also use different $c_w$ to define multiple customer classes in the pooling system. The grand coalition may not be the most profitable coalition under FCFS discipline, because of the increased expected waiting time for some service providers. Thus, the coalition formation should be investigated in the service pooling game.

As for the work of Chapter 2, co-opetition is one interesting extensible strategy. Customers of a service provider would not be attracted by other providers which are members in the coalition under the cooperative point of view. In a co-opetition setting, it is reasonable to define the service arrival rate as a function of the service performance, i.e., the expected waiting time in the system. The co-opetitive strategies could be deduced from two consecutive problems: 1) how to define the customer arrival rates in function of system performances? and 2) the pooling strategy is profitable or not?

Concerning the work of Chapter 3, it would be interesting to develop structural results for the pooling game in order to develop further generic guidelines and insights. It is important to further study the pooling strategy with multiple customer classes, which may model for instance a situation with multiple hospital patient types. The service times, which is presented as the treatment times, are different in different patient

conditions. The characteristics of customer classes could also be described by different patience times, abandonment costs or waiting costs.

In addition to complex operations and human factors, services are also characterised by a high impact of advanced technologies. There are many interesting research questions related to the assessment of the impact of new technologies on customer behavior. For instance, smart phone apps with delay information, multichannel communication (email, chat...) with customers. It would be interesting to study the impact of these features on collaboration strategies.

# Appendix

## Appendix of Chapter 2

### Appendix A

**Further numerical examples for Section 2.5.4**

Further cases related to Table 2.6 are considered here. We vary the number of service companies from 6 to 15 by adding new service providers in the order of the company numbers. We study the impact of coalition growing.

Figures F.1 and F.2 provide the cost allocations obtained by the two allocation rules. We observe that the two allocation methods provide similar solutions. The proportional allocation shares the pooling cost depending on $\lambda_i$, which is appropriate for an individual provider. Also, the $C_d^*$ reduction depends on $\lambda_i$ (Lemma 2.3). Since the Shapley value is calculated as a function of the contributions of each company, this leads to a similarity between the two allocations.

The impact of a new participant on this gap is illustrated in Figure F.3. The gap (absolute difference) between the two allocations for the initial coalition is defined as $AD = \sum_{i=1}^{6} |\varphi_i^{p,\lambda} - sh_i^{opt}|$, for all cases. We find that $AD$ decreases when a new company joins the coalition for the first three cases. The larger the size of the new company is, the higher is the reduction of $AD$. Thus, $sh^{opt}$ becomes more and more similar to $\varphi^{p,\lambda}$ with scale expansion of the coalition. For the last two cases, we select the initial coalition $S_6$, which consists of the relatively large or small companies in the set of $N = \{1, \cdots, 15\}$.

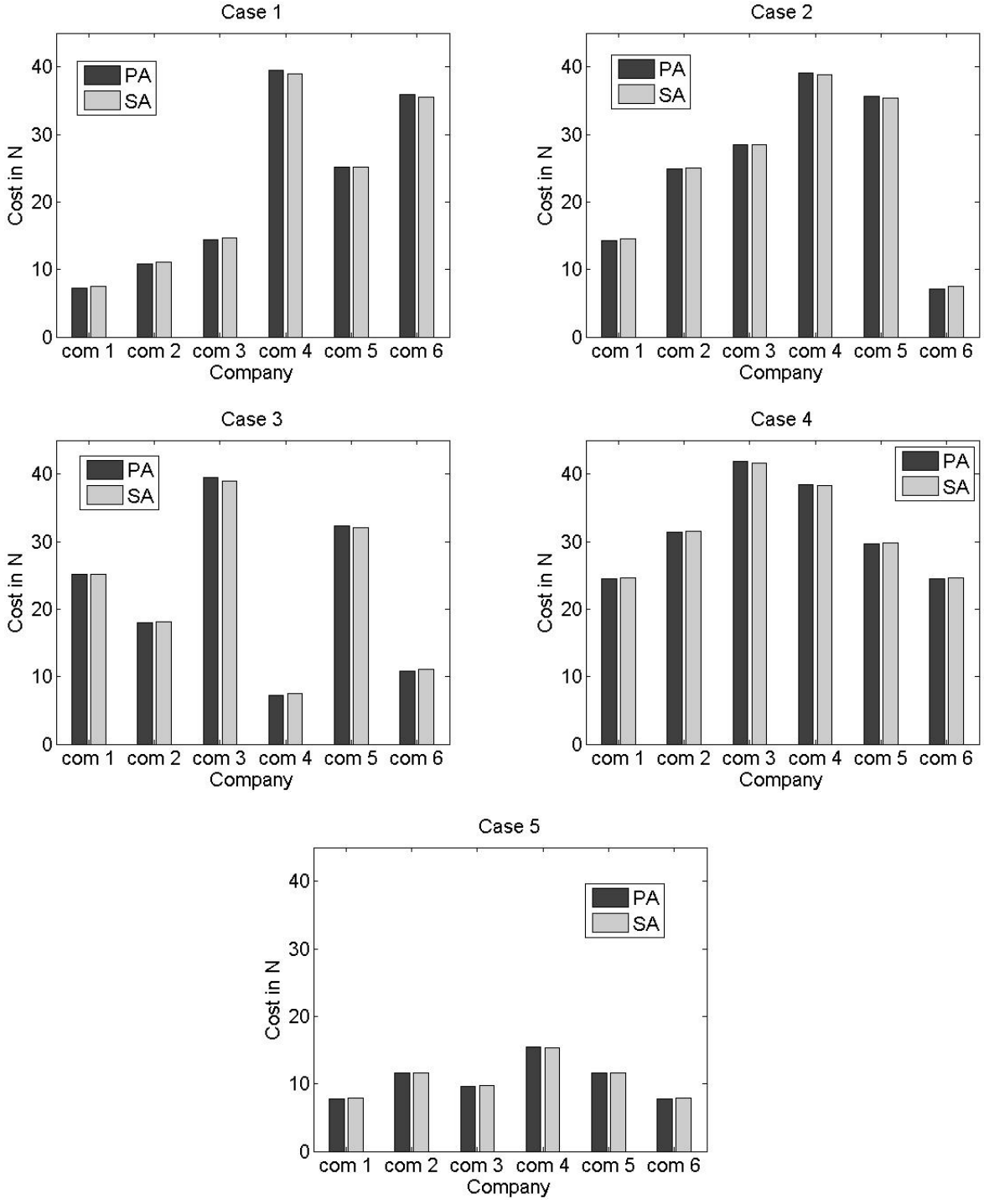



**Figure F.1:** *Proportional allocation cost and Shapley value for the game* $(N, C_{opt})$ *of 6 players with* $f = 0.2$

We find that the joining of a small company increases $AD$ for the coalition of relatively large companies, and the opposite is also true. However, $AD$ decreases or the increase

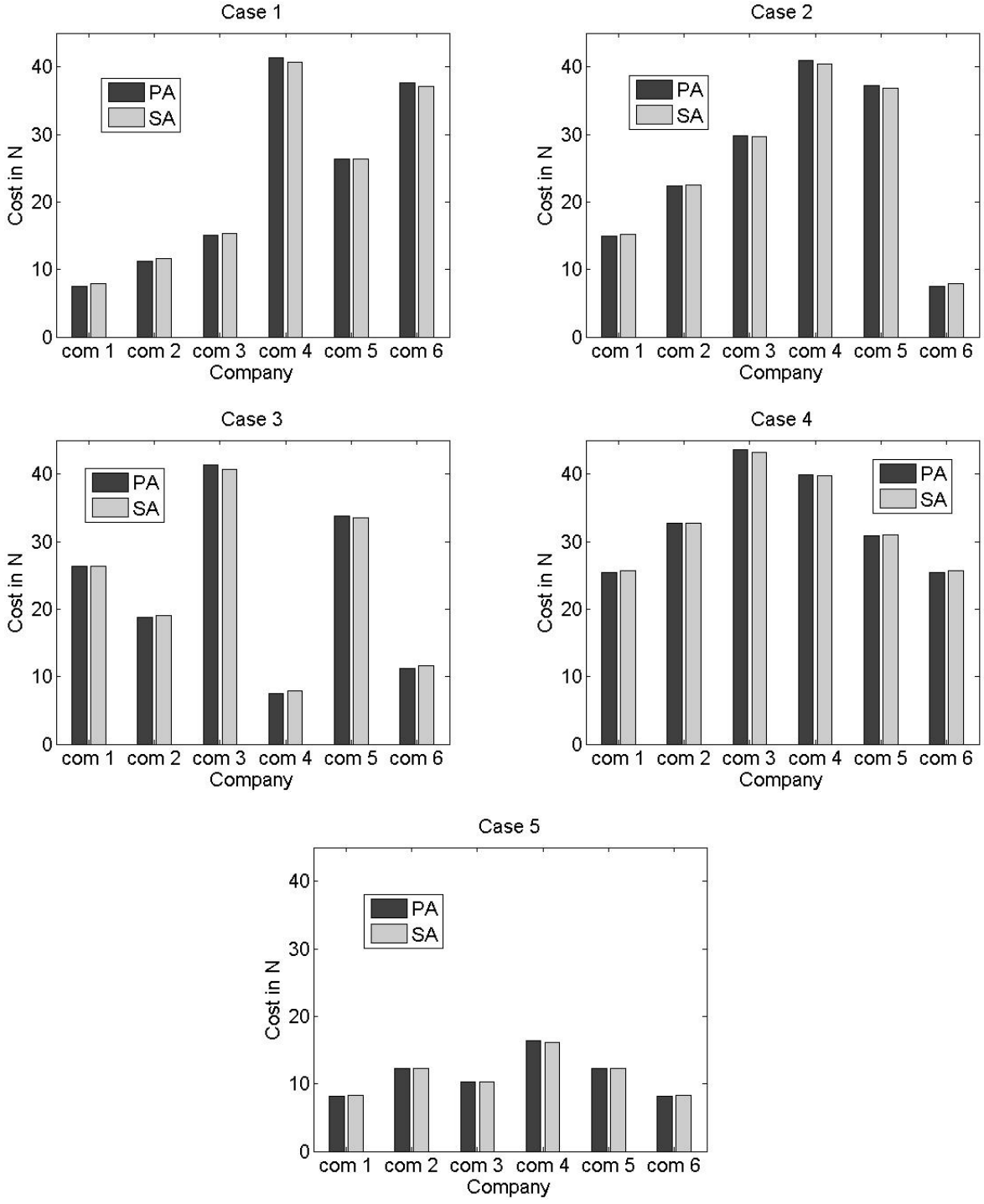



**Figure F.2:** *Proportional allocation cost and Shapley value for the game* $(N, C_{opt})$ *of 6 players with* $f = 1$

will reduce when enough companies join the coalition, and $AD$ is relatively small for the distributed operating costs of $S_6$ ($AD / \sum_{i \in S_6} \varphi_i^{p,\lambda} = 0.5 - 2.5\%$ for all 5 cases).

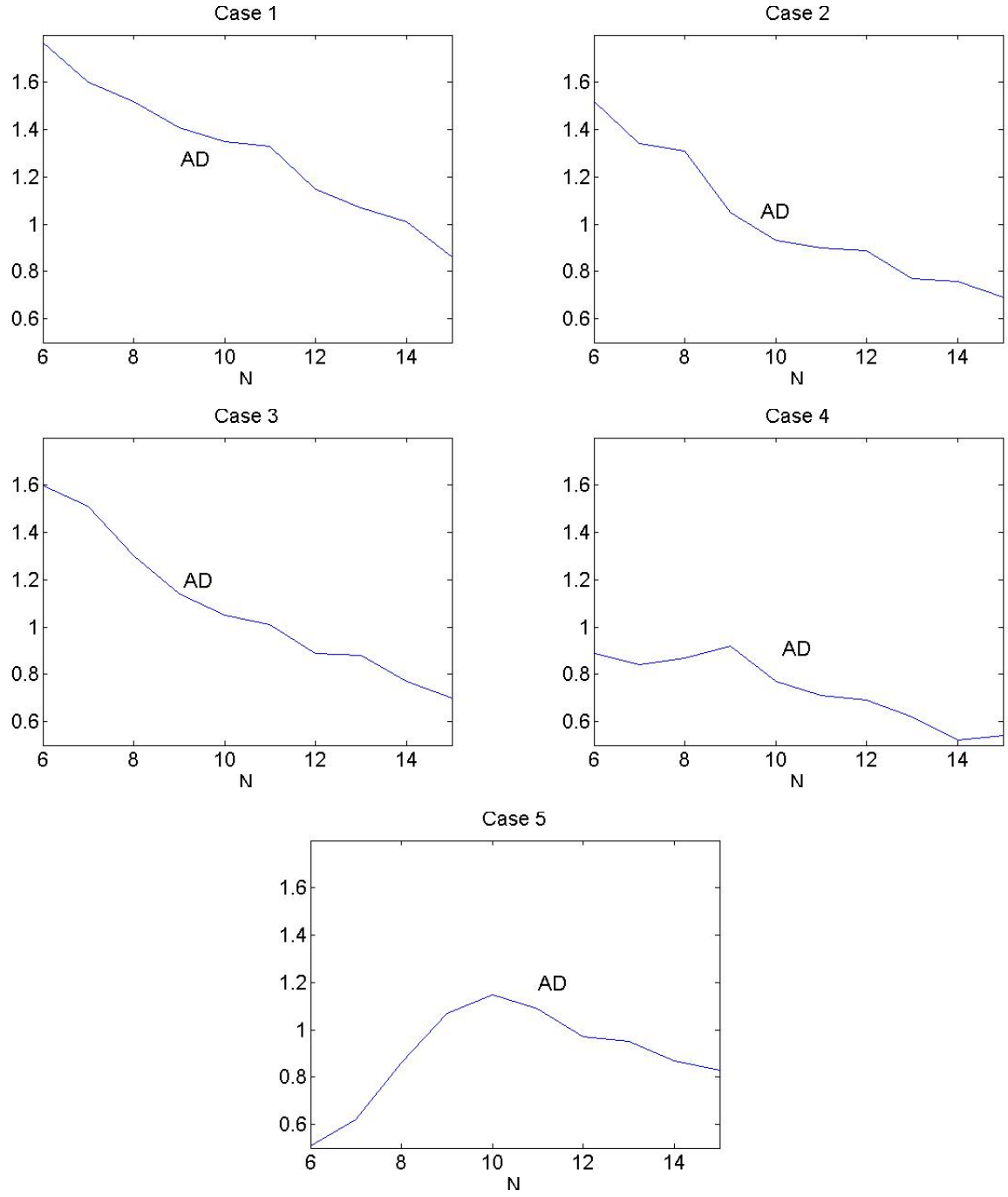



**Figure F.3:** *Difference between the two allocations for coalition* $S_6 = \{1, 2 \ldots 6\}$ *with* $f = 0.2$

Note that the observations agree with the analytical comparison as given in Section 2.5.4.

# Appendix of Chapter 3

## Appendix B

### Numerical illustrations with impatience

For cases with more than 3 players, we consider three typical examples as in [Peng et al., 2015], with low, high and very different offered loads. From Table F.1, we observe that the interval $\Theta \neq \emptyset$ also exists. By comparing the offered loads $\rho_0(i) = \lambda_i/\mu_i$ and $\Theta$, it can be seen that the two bounds of $\Theta$ are related to the coefficient of variation (c.v.) of $\rho_0$.

| System data | | | | Customer arrival rates | | | | | | Offered load $\rho_0$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\mu$ | $\theta^{low}$ | $\theta^{up}$ | **No.** | 1 | 2 | 3 | 4 | 5 | 6 | mean | c.v. |
| 10 | 0.56 | 5769.86 | **1** | 2 | 3 | 3 | 4 | 3.5 | 4 | 0.3250 | 0.2333 |
| - | 0.04 | 32048.84 | **2** | 7 | 9 | 8 | 7 | 8.5 | 9.5 | 0.8167 | 0.1265 |
| - | 1.83 | 1584.23 | **3** | 2 | 7 | 4 | 7.5 | 9 | 3 | 0.5416 | 0.5170 |

**Table F.1:** *System parameters and offered loads of 6 players games*

For a more detailed analysis of $\Theta$, we randomly choose the values of the two variables $\lambda_i$ and $\mu_i$ in $(0, 10)$ (without the constraint of $\lambda_i < \mu_i$) and search the stable interval of $sh^{fix}$ with $n \in \{3, ..., 10\}$. In the 3 players games as shown in Table F.2, it is interesting to see that the interval width $\Theta$ is narrowed with the increase of the coefficient of variation (c.v.) of $\rho_0$ for the cases (fourth to tenth lines) in which $sh^{fix}$ is not stable in $\theta = 0$ and $\theta \to +\infty$ situations.

We test 1000 random cases of $n$ players games for 10 times, respectively with $n \in \{3, ..., 10\}$. Again, we randomly choose the values of $\lambda_i$ and $\mu_i$ in $(0, 10)$ for each random event. From the numerical results, we find out some cases with $|\Theta| = 0$, i.e., there is no $\theta \in \mathbb{R}^+$ with $sh^{fix}$ staying in the core. As shown in Fig. F.4, the probability of $|\Theta| = 0$ in 1000 cases is negligible for $n = 3$ (it is $0.8 - 1.4\%$) and it increases in $n$. When $n = 10$, this probability reaches $5.9 - 8.4\%$.

| Stable range | | | Customer arrival rates | | | Service rates | | | Offered load | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\theta^+$ | $\theta^-$ | **No.** | 1 | 2 | 3 | 1 | 2 | 3 | mean | c.v. |
| 0 | $+\infty$ | **1** | 2.5108 | 6.1604 | 4.7329 | 3.5166 | 8.3083 | 5.8526 | 0.7547 | 0.0646 |
| 0.06 | $+\infty$ | **2** | 5.4972 | 9.1719 | 2.8584 | 7.572 | 7.5373 | 3.8045 | 0.8981 | 0.3078 |
| 0.10 | $+\infty$ | **3** | 6.7970 | 6.5510 | 1.6261 | 9.3745 | 5.1336 | 2.4090 | 0.8921 | 0.3739 |
| 0.33 | 13425.73 | **4** | 3.8156 | 7.6552 | 7.952 | 1.8687 | 4.8976 | 4.4559 | 1.7965 | 0.1334 |
| 0.16 | 6339.85 | **5** | 9.3399 | 6.7874 | 7.5774 | 7.4313 | 3.9223 | 6.5548 | 1.3811 | 0.2221 |
| 0.07 | 1806.43 | **6** | 6.4631 | 7.0936 | 7.5469 | 2.7603 | 6.797 | 6.551 | 1.5124 | 0.4761 |
| 0.66 | 560.05 | **7** | 2.2381 | 7.5127 | 2.551 | 5.0596 | 6.9908 | 8.9090 | 0.6011 | 0.6945 |
| 1.04 | 134.64 | **8** | 1.6261 | 1.19 | 4.9836 | 9.5974 | 3.4039 | 5.8527 | 0.4568 | 0.7737 |
| 0.01 | 112.96 | **9** | 2.5428 | 8.1428 | 2.4352 | 9.2926 | 3.4998 | 1.966 | 1.2796 | 0.8027 |
| 0.47 | 46.57 | **10** | 9.5929 | 5.4721 | 1.3862 | 1.4929 | 2.5750 | 8.4071 | 2.9052 | 1.1023 |

**Table F.2:** *Shapley value's stable range of θ for the system with 3 players games*

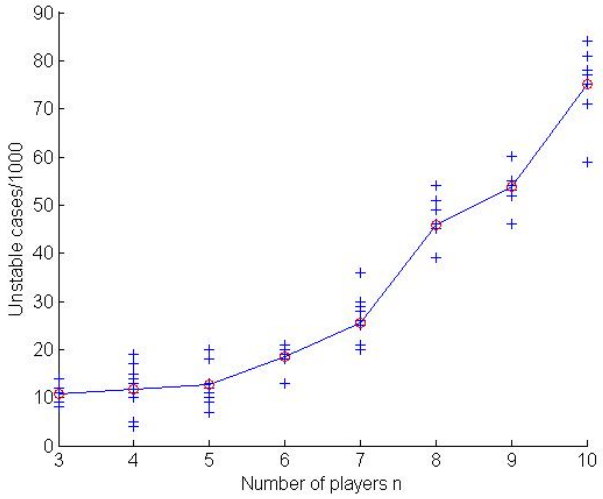



**Figure F.4:** *Number of the cases with* $|\Theta| = 0$ *in 1000 stochastic cases for* $n = \{3, ..., 10\}$

# Appendix C

### Proof of Theorem 3.2

This appendix is devoted to proving Theorem 3.2. Let us first give the following two observations.

**Observation F.1.** For $a_i, b_i \in \mathbb{R}^+$ ($i \in \{1,2,3,4\}$), assume that $a_1 + a_4 = a_2 + a_3$ with $0 \leq a_1 \leq a_2 \leq a_3 \leq a_4$, and $b_1 + b_4 \leq b_2 + b_3$ with $b_1 \geq max\{b_2, b_3, b_4\}$. Then, $a_1b_1 + a_4b_4 \leq a_2b_2 + a_3b_3$.

*Proof.* We define $\widetilde{b}_4 = b_2 + b_3 - b_1$, such that $b_1 + b_4 \leq b_1 + \widetilde{b}_4 = b_2 + b_3$. Thus, $\widetilde{b}_4 \geq b_4$. This implies $a_1b_1 + a_4b_4 \leq a_1b_1 + a_4\widetilde{b}_4 \leq a_2b_1 + a_3\widetilde{b}_4 \leq a_2b_2 + a_3b_3$, which finishes the proof of the observation. □

**Observation F.2.** For $a_i, b_i \in \mathbb{R}^+$ ($i \in \{1,2,3,4\}$), assume that $a_1 + a_4 = a_2 + a_3$ with $0 \leq a_1 \leq a_2 \leq a_3 \leq a_4$, and $b_1 + b_4 \leq b_2 + b_3$ with $b_1 \geq max\{b_2, b_3, b_4\}$. If $a_1b_1 \geq max\{a_2b_2, a_3b_3, a_4b_4\}$, then $a_4b_4 \leq min\{a_1b_1, a_2b_2, a_3b_3\}$ and vice versa.

*Proof.* We define $\widetilde{b}_4 = b_2 + b_3 - b_1 \geq b_4$. We assume that $a_2 - a_1 = a$, $a_3 - a_2 = \triangle a$ and $b_1 - max\{b_2, b_3\} = b$, $\mid b_2 - b_3 \mid = \triangle b$. If $b_2 \geq b_3$, we have

$$
\begin{aligned}
a_2b_2 &= a_1b_1 + A, \\
a_3b_3 &= a_1b_1 + A + B, \\
a_4b_4 \leq a_4\widetilde{b}_4 &= a_1b_1 + 2A + B - C,
\end{aligned}
$$

with $A = aa_1 - \triangle aa_1 - ab$, $B = \triangle a(a_1 - a) - \triangle b(b_1 + b) - \triangle a \triangle b$ and $C = 2ab + \triangle aa + \triangle bb$. Since $a_1b_1 \geq max\{a_2b_2, a_3b_3, a_4b_4\}$, we obtain $A \leq 0$, $A + B \leq 0$ and $C \geq 0$. Thus, $a_4b_4 \leq min\{a_1b_1, a_2b_2, a_3b_3\}$. Similarly, we could prove the same result in the case of $b_2 < b_3$. When $a_4b_4 \geq max\{a_1b_1, a_2b_2, a_3b_3\}$, we define $\widetilde{b}_1 = b_2 + b_3 - b_4$ such that $\widetilde{b}_1 \geq b_1$. We then obtain $a_1b_1 \leq min\{a_2b_2, a_3b_3, a_4b_4\}$ from the symmetry of $a_i$ and $b_i$, which completes the proof of this observation. □

We now proceed to the concavity proof of the queue length.

*Proof.* of **Theorem 2.** We define the same $Y$, $X(\theta)$ and $X^+(\theta)$ as in the proof of Lemma 3.1. The decreasing property in $\theta \in (0, +\inf)$ is easy to get from the transition rate of $Y$, $\lambda/(\mu + j\theta)$, which is decreasing in $\theta$. Following similar arguments as for Lemma 3.1, we can prove this theorem by proving that $X$ is SDCX(sp) in $\theta$ if $\theta \leq \mu$. In this proof, we

focus on the SDCX(sp) of $X$ in $\theta$, using a similar method used for that of $X$ in $\mu$ in the proof of Theorem 1 in [Armony et al., 2009].

We use the definition of sample path convexity as defined in [Shaked and Shanthikumar, 1988]. For initialization, we choose four abandonment rates $0 \leq \theta_1 \leq \theta_2 \leq \theta_3 \leq \theta_4 \leq \mu$ with $\theta_1 + \theta_4 = \theta_2 + \theta_3$. All the other system parameters $\lambda$ and $\mu$ are held constant. If $X$ is SDCX(sp), then $X_i =_{st} X(\theta_i)$, $i \in \{1,2,3,4\}$, defined on the same probability space as the uniformized Markov process of $X_i$ defined at time $m$, satisfies

Condition 1. $X_1(m) + X_4(m) \geq X_2(m) + X_3(m), a.s.$

Condition 2. $X_1(m) \geq max\{X_2(m), X_3(m), X_4(m)\}, a.s.,$ for all $m \in \mathbf{N}$.

For uniformization, we define $K\theta_i$ with $K \in \mathbf{N}^+$ as the upper limit of the abandonment rate, which means that there are at maximum $K$ customers that has the intention to abandon the queue at a given time. The maximum transition rate of all $X_i$ is upper bounded by $v = \lambda + \mu + K\theta_4$. Thus, we define $X_i^K$ as the uniformized version of $X_i$. Now, we will show that for any $K$, Conditions 1 and 2 hold for any $m \in \mathbf{N}$, then $X^K$ is SDCX(sp). To simplify the presentation, we write $X_i^K$ as $X_i$ in the following arguments.

The proof of the two conditions is done by induction on the discrete time $m$. Firstly, we suppose that the equality of Condition 1, denoted by $\widetilde{1}$, and Condition 2 hold at time $m$. We next build the transition probability to satisfy Conditions 1 and 2 at time $m+1$.

With a probability of $\lambda/v$, new customers arrive into all the four systems. The similar transition definition is used for the service completion, services are completed simultaneously in all systems with a probability of $\mu/v$. Consider a special definition in case of $X_4 = 0$, which will be discussed later on.

For the abandonment process, there is a probability of $min\{(X_i - 1)^+, K\}\theta_i/v$ that one customer abandons from system $i$, $i \in \{1,2,3,4\}$. Define $X_i^{(K)+} = min\{(X_i - 1)^+, K\}$. From the concavity of $f = min\{x, K\}$, we have $X_1^{(K)+} + X_4^{(K)+} \leq X_2^{(K)+} + X_3^{(K)+}$ for the not empty queues ($X_i \neq 0$ for all systems). To further specify the abandonment transition probability for each system, we define four cases depending on the order of

$X_i^{(K)+}\theta_i$ denoted by $\vartheta_i$. From Observation F.1, Conditions $\widetilde{1}$ and 2, it follows

$$\vartheta_1 + \vartheta_4 \leq \vartheta_2 + \vartheta_3, \tag{F.1}$$

at time $m$.

Consider the existence of empty queues ($X_i = 0$ for some systems). According to Conditions $\widetilde{1}$ and 2, there are four possible situations: no customer in Systems $\{1,2,3,4\}$, $\{2,4\}$, $\{3,4\}$ or 4. When Systems $\{1,2,3,4\}$, $\{2,4\}$ or $\{3,4\}$ are empty, Inequality (F.1) still holds. The only exception happens when System 4 is the only empty system. In this case, we have $\vartheta_1 + \mu \leq (\vartheta_2 + \mu) + (\vartheta_3 + \mu)$ with $\theta_4 \leq \mu$. Here, we deal with the service completion and the customer abandonment process together and choose $\vartheta_i$ equal to $\vartheta_i + \mu$. Thus, Inequality (F.1) holds with $\vartheta_4 = \vartheta_{(4)} = 0$.

From Observation F.2, the four following cases meet Inequality (F.1). We define $\vartheta_{(i)}$ for $i \in \{1,2,3,4\}$, such that $0 \leq \vartheta_{(4)} \leq \vartheta_{(3)} \leq \vartheta_{(2)} \leq \vartheta_{(1)}$. Let a random variable $p_a \sim Uniform(0,1)$.

**Case 1.** If $\vartheta_{(1)} + \vartheta_{(4)} \leq \vartheta_{(2)} + \vartheta_{(3)}$. Let $\widetilde{\vartheta} = max\{K\theta_4, \vartheta_{(2)} + \vartheta_{(3)} - \vartheta_{(4)}\}$.

- I) If $p_a \leq \vartheta_{(4)}/\widetilde{\vartheta}$, there are customers abandoning from all systems;
- II) If $\vartheta_{(4)}/\widetilde{\vartheta} \leq p_a \leq \vartheta_{(3)}/\widetilde{\vartheta}$, there are customers abandoning from Systems $\{1,3\}$;
- III) If $\vartheta_{(3)}/\widetilde{\vartheta} \leq p_a \leq \vartheta_{(1)}/\widetilde{\vartheta}$, there are customers abandoning from Systems $\{1,2\}$;
- IV) If $\vartheta_{(1)}/\widetilde{\vartheta} \leq p_a \leq (\vartheta_{(2)} + \vartheta_{(3)} - \vartheta_{(4)})/\widetilde{\vartheta}$, there is one customer abandoning from System 2.

**Case 2.** If $\vartheta_{(2)} + \vartheta_{(3)} \leq \vartheta_{(1)} + \vartheta_{(4)}$. Let $\widetilde{\vartheta} = max\{K\theta_4, \vartheta_{(1)}\}$.

- I) If $p_a \leq \vartheta_{(4)}/\widetilde{\vartheta}$, there are customers abandoning from all systems;
- II) If $\vartheta_{(4)}/\widetilde{\vartheta} \leq p_a \leq \vartheta_{(3)}/\widetilde{\vartheta}$, there are customers abandoning from Systems $\{1,3\}$;
- III) If $\vartheta_{(3)}/\widetilde{\vartheta} \leq p_a \leq (\vartheta_{(2)} + \vartheta_{(3)} - \vartheta_{(4)})/\widetilde{\vartheta}$, there are customers abandoning from System $\{1,2\}$;
- IV) If $(\vartheta_{(2)} + \vartheta_{(3)} - \vartheta_{(4)})/\widetilde{\vartheta} \leq p_a \leq \vartheta_{(1)}/\widetilde{\vartheta}$, there is one customer abandoning from System 1.

**Case 3.** If $\vartheta_{(2)} + \vartheta_{(4)} \leq \vartheta_{(1)} + \vartheta_{(3)}$. Let $\widetilde{\vartheta} = max\{K\theta_4, \vartheta_{(1)}\}$.

- I) If $p_a \leq \vartheta_{(4)}/\widetilde{\vartheta}$, there are customers abandoning from all systems;
- II) If $\vartheta_{(4)}/\widetilde{\vartheta} \leq p_a \leq \vartheta_{(3)}/\widetilde{\vartheta}$, there are customers abandoning from Systems $\{1,2,3\}$;
- III) If $\vartheta_{(3)}/\widetilde{\vartheta} \leq p_a \leq \vartheta_{(2)}/\widetilde{\vartheta}$, there are customers abandoning from Systems $\{1,2\}$;
- IV) If $\vartheta_{(2)}/\widetilde{\vartheta} \leq p_a \leq \vartheta_{(1)}/\widetilde{\vartheta}$, there is one customer abandoning from System 1;

**Case 4.** If $\vartheta_{(3)} + \vartheta_{(4)} \leq \vartheta_{(1)} + \vartheta_{(2)}$. Let $\widetilde{\vartheta} = max\{K\theta_4, \vartheta_{(1)} + \vartheta_{(2)} - \vartheta_{(3)}\}$.

- I) If $p_a \leq \vartheta_{(4)}/\widetilde{\vartheta}$, there are customers abandoning from all systems;
- II) If $\vartheta_{(4)}/\widetilde{\vartheta} \leq p_a \leq \vartheta_{(3)}/\widetilde{\vartheta}$, there are customers abandoning from Systems $\{1,2,3\}$;
- III) If $\vartheta_{(3)}/\widetilde{\vartheta} \leq p_a \leq \vartheta_{(1)}/\widetilde{\vartheta}$, there is one customer abandoning from System 1;
- IV) If $\vartheta_{(1)}/\widetilde{\vartheta} \leq p_a \leq (\vartheta_{(1)} + \vartheta_{(2)} - \vartheta_{(3)})/\widetilde{\vartheta}$, there is one customer abandoning from System 2;

We next show that if Conditions $\widetilde{1}$ and 2 hold at time $m$, then Conditions 1 and 2 hold at time $m+1$.

For the arrival process, it is obvious that Conditions 1 and 2 hold at $m+1$. From Condition 2, service completions from the left hand side of Condition 1 are not higher than those from the other hand side, and $X_1$ stays as the longest queue. This is because service completions 1) happen in both systems if $X_2 = X_1$ or $X_3 = X_1$; 2) happen in the three systems or all systems if $X_2 = X_3 = X_1$; 3) happen in all systems if all queue lengths are identical. For the abandonment process, it is clear that the number of abandoned customers in Systems $\{1,4\}$ is not higher than that in Systems $\{2,3\}$ after one step transition in the case of $X_i^{(K)+} > 0$, i.e., Inequality (F.1) holds for all the four cases. Now, we discuss special situations in the customer abandonment process.

- I) All the four queues $X_i^{(K)+}$ are empty, there is no abandonment in all systems;
- II) If two queues ($\{2,4\}$ or $\{3,4\}$) are empty, we have $\vartheta_{(2)} \leq \vartheta_{(1)}$ corresponding to Case 2. System 1 has an abandonment only if System 2 does;
- III) If $X_4^+ = 0$, System 1 has an abandonment when at least one customer abandons in the System 2 or 3 (in Case 1, 3 or 4);

Thus, Condition 1 holds in the abandonment process at time $m+1$. For Condition 2, we have

- I) If $X_1(m) > max\{X_2(m), X_3(m), X_4(m)\}$, it is obvious that Condition 2 holds at time $m+1$;
- II) If $\vartheta_1 = \vartheta_{(4)}$, a customer abandons from System 1 when customers have abandoned from all the other 3 systems;
- III) If $\vartheta_1 = \vartheta_{(3)}$ with $X_1 = X_2 > X_4$ (or $X_1 = X_3 > X_4$) and $X_3 = X_4$ (or $X_2 = X_4$), we have $\vartheta_2 \geq \vartheta_1$ and $\vartheta_4 \geq \vartheta_3$. Then, $\vartheta_4 \neq \vartheta_{(4)}$. The corresponding case is Case 2. and we have $\vartheta_2 = \vartheta_{(1)}$. Customers abandon from Systems $\{1,2\}$ together;
- IV) If $\vartheta_1 = \vartheta_{(2)}$ with $X_1 = X_2 > X_4$ (or $X_1 = X_3 > X_4$), we have $\vartheta_2 = \vartheta_{(1)} \geq \vartheta_1$ . In the corresponding cases (Cases 2 and 3), a customer abandons from System 1 only if a customer has abandoned from System 2;
- V) If $X_1(m) = X_2(m) = X_3(m) = X_4(m)$, we have $0 \leq \vartheta_1 \leq \vartheta_2 \leq \vartheta_3 \leq \vartheta_4$, which coincides with Case 1 and $X_1 = \vartheta_{(4)}$.

As summary in the above paragraphs, it is shown that if Conditions $\widetilde{1}$ and 2 hold at time $m$, Conditions 1 and 2 hold at time $m+1$. If strict inequality of Condition 1 holds, we define $\widetilde{X}_4(m) = max\{0, X_2(m) + X_3(m) - X_1(m)\} \leq X_4(m)$ and $\widetilde{X}_1(m) = min\{X_2(m) + X_3(m), X_1(m)\} \leq X_1(m)$. It means that if $X_2 + X_3 - X_1 \geq 0$, we decrease $\widetilde{X}_4 = X_2 + X_3 - X_1 \geq 0$ and keep $\widetilde{X}_1 = X_1$. We have $\widetilde{X}_1 = X_2 + X_3$ and $\widetilde{X}_4 = 0$ in the opposite case. We have $\widetilde{X}_1(m) + \widetilde{X}_4(m) = X_2(m) + X_3(m)$. Then, Conditions 1 and 2 hold for $\{\widetilde{X}_1, X_2, X_3, \widetilde{X}_4\}$ at time $m+1$. For each $\theta_i$, we denote the complementary cumulative distribution function (ccdf) by $\overline{F}_{\theta_i}(y;x) = P_{\theta_i}\{X(m+1) > y | X(m) = x\}$ and its inverse $\overline{F}_{\theta_i}(v;x)^{-1} = inf\{y : \overline{F}_{\theta_i}(y;x) \leq v\}$, $v \in [0,1]$. Define $X_i(m+1) = \overline{F}_{\theta_i}^{-1}(\overline{F}_{\theta_i}(\widetilde{X}_i(m+1); \widetilde{X}_i(m)); X_i(m))$ for $i = \{1,4\}$. From the transition probabilities determined in the previous paragraphs, it follows that $\overline{F}_{\theta_i}(y;x)$ is non-decreasing in $x$. Thus, $(X_i(m+1) - X_i(m)) \geq (\widetilde{X}_i(m+1) - \widetilde{X}_i(m))$ for $i = \{1,4\}$ and Conditions 1 and 2 hold for $\{X_1, X_2, X_3, X_4\}$ at time $m+1$ with the corresponding transition probabilities. The proof of Conditions 1 and 2 for $X^K$ is now completed. Therefore, $\mathbb{E}(X^K)$ is decreasing and convex in $\theta$. Since $X^K(\theta) \xrightarrow[K \longrightarrow \infty]{st} X(\theta)$ for each $\theta \in [0, \mu]$, $\mathbb{E}(X)$ is decreasing and convex in $\theta$ from Proposition 2.11 in [Shaked and Shanthikumar, 1988]. The proof of the theorem is completed.

□

# Appendix D Résumé étendu

## 1. Introduction & motivation

Le secteur des services est devenu le secteur le plus important en nombre d'emplois occupés dans l'économie mondiale (figure 1.1, page 6), en particulier dans les pays développés, les services représentent jusqu'à 70% de la production nationale (PIB) et sont devenus leur principal moteur de croissance économique. Par exemples, quatre sur cinque emplois aux Etats-Unis sont fournis par le secteur de services ; le secteur tertiaire français occupait 76,8% de la population active en 2015.

Beaucoup de nos activités quotidiennes dépendent des services et des fournisseurs de services, de l'e-mail que nous vérifions le matin au service de transport public que nous prenons pour aller à notre lieu de travail, du restaurant dans lequel nous déjeunons à midi au colis que nous recevons pendant la journée. Les services sont partout dans notre vie, y compris dans les domaines de la finance (banques, stocks), la santé (médecin personnel, hôpital), la communication (courrier électronique, réseau 4G), les services publics (électricité, police), etc. [Daskin, 2010]

Dans le contexte de la mondialisation économique, la concurrence et la coopération dans les industries de services sont devenues de plus en plus communes: la concurrence des prix entre les chaînes de restauration rapide, le regroupement des entreprises de télécommunication, les services collaboratifs d'après-vente et de maintenance dans l'industrie électronique, pour n'en citer que quelques-uns. Dans cette thèse, nous étudions des stratégies collaboratives dans les systèmes de service homogènes. Nous nous concentrons en particulier sur les stratégies de pooling des ressources. Notre approche consiste à utiliser la modélisation des systèmes de service par les files d'attents et la théorie des jeux pour l'analyse des interactions entre les fournisseurs de services. Dans ce qui suit, nous discutons brièvement les stratégies de collaboration et de pooling des ressources.

**Stratégies collaboratives dans les services**

Afin d'améliorer les performances du système ou de réduire les dépenses, il existe plusieurs méthodes coopératives basiques: la coopération entre les files d'attente, par exemple, la dispatching d'arrivée simultanées entre les agences ou le réacheminement entre des serveurs différents [Katta and Sethuraman, 2006, Kayi and Ramaekers, 2010]; le pooling des services, p.ex., le pooling des capacités de service ou la répartition en personnel [Guo et al., 2013]; le cross-training [Tekin et al., 2014]; la collaboration avec des fournisseurs tiers, par exemple l'externalisation de services [Aksin et al., 2008], etc. Il est parfois utile de combiner ces méthodes pour aboutir à une structure collaborative plus rentable [Anily and Haviv, 2014].

Les méthodes de coopération pourrait être classées sous trois formes typiques (figure 1.2, page 8): la forme verticale, les collaborations entre les différents éléments du système, par exemple, les forfaits de téléphone portable signé par les clients avec les opérateurs télécoms, les cartes annuelles d'accès au club de sport; la forme horizontale, la collaboration entre les serveurs homogènes, par ex., les services après-vente de produits électroniques de différentes marques; et les externalisations, la collaboration avec une partie tiers aux systèmes de service, par exemple, le service client sous-traité à l'étranger.

Parmi la majorité des gains réalisée par les activités collaboratives, la réduction des coûts est le facteur le plus important pour les fournisseurs de services. Le coût de la capacité de service et le coût d'attente des clients dans la file d'attente ou le système sont largement utilisés dans la littérature [Anily and Haviv, 2010, Özen et al., 2011, Karsten et al., 2015b, Yu et al., 2015].

**Pooling des ressources**

Depuis la première étude de [Stidham, 1970], le pooling des systèmes de file d'attente a été largement étudié dans la littérature pour la conception des systèmes de service.

Il est bien connu que le pooling de la capacité de service amène naturellement à des économies d'échelle dans les études de gestion opérationnelle [Smith and Whitt, 1981, Bell and Williams, 2005]. Cette amélioration de l'efficacité opérationnelle est générée par la disparition des ressources de service libres dans le système quand il est en présence de congestion dans les files d'attente. Ceci est valide à la fois dans le cas de départements d'une même entité économique, comme par exemple, le pooling des réservations dans un restaurant [Thompson and Kwortnik, 2008], ou dans le cas de plusieurs entités indépendantes [González and Herrero, 2004, Garcia-Sanz et al., 2008, Anily and Haviv, 2010, Kayi and Ramaekers, 2010, Tekin et al., 2014, Anily and Haviv, 2014].

Les applications en pratique pour le pooling des services parmi les fournisseurs de services homogènes sont nombreuses. Par exemple, les différents départements d'un hôpital peuvent partager une salle d'opération commune afin de réduire les dépenses. Ils pourraient également partager leur capacité en termes de lits dans les chambres d'hôpital, ce qui permettrait de soulager la congestion. Un autre exemple est dans le contexte de service après-vente pour certaines catégories de produits électroniques. Ces produits sont susceptibles d'avoir un faible taux de demande après-vente pour chaque détaillant individuellement. En conséquence, les détaillants pourraient fournir un service après-vente conjoint pour réduire les coûts de démarrage du service et améliorer la qualité du service. Pour les services d'aviation, le service d'enregistrement en commun pour différentes compagnies aériennes est un exemple supplémentaire pour les applications du pooling d'ensemble services.

Parmi les perspectives de recherche, on peut définir les stratégies de pooling des ressources plus précisé. Selon les participants, on peut distinguer les cas des différentes branches d'une même entité économique [Alptekinoğlu et al., 2013] ou de plusieurs entités indépendantes [Anily and Haviv, 2014]; Selon les méthodes de pooling, on peut distinguer les cas avec pooling des ressources partiel [Chao et al., 2003] ou complète [Anily and Haviv, 2010]; Selon la définition de la valeur des systèmes issue de la théorie des jeux, on peut distinguer les cas avec la valeur transférable ou la valeur non-transférable, par exemples, le coût, est une valeur transférable [Karsten et al., 2015b], et la réputation

de fournisseur est non-transférable [Toivonen, 2014]. Dans cette thèse, nous travaillons sur les stratégies du pooling des ressources complètes entre plusieurs entités indépendantes, et la valeur des systèmes est mesurée par les coûts globaux.

Il existe des similitudes entre le management opérationnelle de service et de biens manufacturés, tous les deux sont concernés par l'efficience, l'efficacité, les problèmes de qualité et motivés par la réduction des coûts. Contrairement à l'existence de travaux de recherche dans l'industrie manufacturière, les recherches relatifs à l'industrie de service ne répendant pas aux exigences de son développement économique énorme.

Les services sont principalement caractérisés par des opérations complexes et un impact élevé des facteurs humains. Dans cette thèse, nous tenons compte de ces deux aspects par l'analyse de l'impact de la variabilité de la durée du service et de l'abandon du client, respectivement. Nous étudions le problème dans lequel les fournisseurs de service indépendants pourraient être amenés à coopérer entre eux. Nous considérons la stratégie de pooling des ressources dans différents systèmes de services et fournissons des stratégies de pooling correspondantes en utilisant la théorie des jeux coopératifs.

## 2. Objectif & contributions

L'objectif de cette thèse est d'étudier l'impact des caractéristiques de la variabilité des services et de l'abandon des clients sur les stratégies de collaboration. Motivés par la réduction des coûts, nous traitons le problème de pooling des ressources entre les fournisseurs de services indépendants. Nous utilisons la théorie des files d'attente pour la modélisation de ces caractéristiques. Plus concrètement, nous posons les deux questions suivantes: 1) comment former les coalitions? Et 2) quelle règle d'allocation doit être choisie pour garantir la stabilité de la coalition? Nous utilisons la théorie des jeux coopératifs, qui fournit des concepts intéressants pour analyser les structures de coalitions profitables et résoudre le problème de partage des coûts entre les participants.

Les principales contributions de cette thèse peuvent être résumées comme suit.

Premièrement, nous étudions le problème du partage des coûts entre les fournisseurs

de services indépendants dans le cadre d'un système de pooling des capacités de service avec des temps de service distribués suivant une loi générale. L'amélioration effective est obtenue en réduisant les ressources inutilisées en cas de congestion. Nous modélisons à la fois le fournisseur de services et la coalition coopérative par des files d'attente à un seul serveur avec des délais de service suivant une loi générale. Pour les deux situations de pooling avec une capacité de service fixe et de pooling avec une capacité de service optimisée, nous définissons les jeux coopératifs correspondants et analysons les allocations du cœur du jeu. Pour le cas avec capacité fixe, nous prouvons que le cœur est non vide. La fonction caractéristique n'est ni concave ni monotone dans le jeu mentionné. Cependant, nous prouvons que le jeu de pooling de service avec la capacité de service optimisée est concave. Pour ce jeu concave, nous trouvons deux règles d'allocation stables et illustrons une stratégie combinée d'allocation des coûts.

Deuxièmement, nous considérons un groupe de fournisseurs de service homogènes et indépendants, où un client quitte le système sans service chaque fois que son attente dans la file d'attente dépasse son seuil de temps de patience. L'avantage de la collaboration dans ces systèmes avec l'abandon des clients, n'est pas seulement le partage de ressources instantanées disponibles, mais aussi la réduction des clients qui abandonnent. Selon les hypothèses markoviennes sur les temps d'arrivée, de service et de patience, nous définissons un jeu coopératif avec utilité transférable et une capacité de service fixe pour chaque individu et chaque coalition. Nous prouvons que la grande coalition est la coalition la plus profitable et que le jeu a un cœur non vide. Nous examinons ensuite l'impact de l'abandon sur la stabilité de la valeur de Shapley. De plus, nous démontrons la concavité de la longueur de la file d'attente en fonction du taux d'abandon et donnons une condition selon laquelle la valeur Shapley est située dans le cœur. Nous étudions également le problème de partage des coûts du jeu coopératif relatif avec la capacité de service optimisée et prouvons que la règle d'allocation proportionnelle selon les taux d'arrivée des clients donne une allocation stable dynamique à tous les sous-jeux relatifs.

Dans les études précédentes, nous utilisons les hypothèses de $\ll$ super-serveur $\gg$. La raison principale de cette hypothèse est que traiter les systèmes des files d'attente

multiserveurs avec les temps de service distribué selon une loi générale et avec abandon des clients est très complexe. Pour évaluer la qualité de ces hypothèses, nous étudions numériquement le problème des stratégies de pooling des services dans le cas multi-serveurs. Bien qu'il soit intuitif de s'attendre à des améliorations d'efficacité dans les systèmes de pooling multiserveurs, il n'est pas évident de conclure que tous les membres bénéficieront du pooling comme c'est le cas pour le $\ll$ super-serveur $\gg$. Nous comparons les deux cadres de pooling et nous évaluons numériquement les effets de la variabilité des services et de l'abandon des clients dans les deux jeux correspondants.

# Bibliography

[Aksin et al., 2008] Aksin, O., de Véricourt, F., and Karaesmen, F. (2008). Call center outsourcing contract analysis and choice. *Management Science*, 54(2):354–368.

[Alptekinoğlu et al., 2013] Alptekinoğlu, A., Banerjee, A., Paul, A., and Jain, N. (2013). Inventory pooling to deliver differentiated service. *Manufacturing & Service Operations Management*, 15(1):33–44.

[Andradóttir et al., 2017] Andradóttir, S., Ayhan, H., and Down, D. (2017). Resource pooling in the presence of failures: Efficiency versus risk. *European Journal of Operational Research*, 256:230–241.

[Anily and Haviv, 2007] Anily, S. and Haviv, M. (2007). The cost allocation problem for the first order interaction joint replenishment model. *Operations Research*, 55(2):292–302.

[Anily and Haviv, 2010] Anily, S. and Haviv, M. (2010). Cooperation in service systems. *Operations Research*, 58(3):660–673.

[Anily and Haviv, 2014] Anily, S. and Haviv, M. (2014). Subadditive and homogeneous of degree one games are totally balanced. *Operation Research*, 62(4):788–793.

[Armony et al., 2009] Armony, M., Plambeck, E., and Seshadri, S. (2009). Sensitivity of optimal capacity to customer impatience in an unobservable M/M/S queue (why you shouldn't shout at the DMV). *Manufacturing and Service Operations Management*, 11(4):19–32.

[Batt and Terwiesch, 2015] Batt, R. and Terwiesch, C. (2015). Waiting patiently: An empirical study of queue abandonment in an emergency department. *Management Science*, 61(1):39–59.

[Bell and Williams, 2005] Bell, S. and Williams, R. (2005). Dynamic scheduling of a parallel server system in heavy traffic with complete resource pooling: Asymptotic optimality of a threshold policy. *Electronic Journal of Probability*, 10(33):1044–1115.

[Benjaafar et al., 2005] Benjaafar, S., Cooper, W., and Kim, J. (2005). On the benefits of pooling in production-inventory systems. *Management Science*, 51(4):548–565.

[Bondareva, 1963] Bondareva, O. (1963). Some applications of linear programming methods to the theory of cooperative games. *Problemy Kibernetiki*, 10:119–139.

[Chao et al., 2003] Chao, X., Liu, L., and Zheng, S. (2003). Resource allocation in multisite service systems with intersite customer flows. *Management Science*, 49(12):1739–1752.

[Chun, 2006a] Chun, Y. (2006a). No-envy in queueing problems. *Economic Theory*, 29(1):151–162.

[Chun, 2006b] Chun, Y. (2006b). A pessimistic approach to the queueing problem. *Mathematical Social Sciences*, 51(2):171–181.

[Cruijssen et al., 2007] Cruijssen, F., Cools, M., and Dullaert, W. (2007). Horizontal cooperation in logistics: opportunities and impediments. *Transportation Research Part E: Logistics and Transportation Review*, 43(2):129–142.

[Dai and He, 2010] Dai, J. and He, S. (2010). Customer abandonment in many-server queues. *Mathematics of Operations Research*, 35(2):347–362.

[Daskin, 2010] Daskin, M. (2010). *Service science*, volume 1. Wiley Online Library.

[Dijk and Sluis, 2008] Dijk, N. and Sluis, E. (2008). To pool or not to pool in call centers. *Production and Operations Management*, 17(3):296–305.

[Drechsel and Kimms, 2010] Drechsel, J. and Kimms, A. (2010). Computing core allocations in cooperative games with an application to cooperative procurement. *International Journal of Production Economics*, 128(1):310–321.

[Driessen, 1988] Driessen, T. S. H. (1988). *Cooperative Games, Solutions and Applications*. Kluwer Academic Publishers.

[Dror et al., 2012] Dror, M., Hartman, B., and Chang, W. (2012). The cost allocation issue in joint replenishment. *International Journal of Production Economics*, 135(1):242–254.

[Elomri et al., 2012] Elomri, A., Ghaffari, A., Jemai, Z., and Dallery, Y. (2012). Coalition formation and cost allocation for joint replenishment systems. *Production and Operations Management*, 21(6):1015–1027.

[Ferguson, 2014] Ferguson, T. (2014). *Mathematical Statistics: A Decision Theoretic Approach*, volume 1. Academic press.

[Fiestras-Janeiro et al., 2013] Fiestras-Janeiro, M., García-Jurado, I., Meca, A., and Mosquera, M. (2013). A new cost allocation rule for inventory transportation systems. *Operations Research Letters*, 41(5):449–453.

[Frisk et al., 2010] Frisk, M., Göthe-Lundgren, M., Jörnsten, K., and Rönnqvist, M. (2010). Cost allocation in collaborative forest transportation. *European Journal of Operational Research*, 205(2):448–458.

[Garcia-Sanz et al., 2008] Garcia-Sanz, M., Fernández, F., Fiestras-Janeiro, M., García-Jurado, I., and Puerto, J. (2008). Cooperation in markovian queueing models. *European Journal of Operational Research*, 188(2):485–495.

[Garnett et al., 2002] Garnett, O., Mandelbaum, A., and Reiman, M. (2002). Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4:208–227.

[Gerchak and Gupta, 1991] Gerchak, Y. and Gupta, D. (1991). On apportioning costs to customers in centralized continuous review inventory systems. *Journal of Operations Management*, 10(4):546–551.

[Gillies, 1959] Gillies, D. (1959). Solutions to general non-zero-sum games. *Contributions to the Theory of Games*, 4(40):47–86.

[González and Herrero, 2004] González, P. and Herrero, C. (2004). Optimal sharing of surgical costs in the presence of queues. *Mathematical Methods of Operations Research*, 59(3):435–446.

[Guajardo and Rönnqvist, 2015] Guajardo, M. and Rönnqvist, M. (2015). Cost allocation in inventory pools of spare parts with service-differentiated demand classes. *International Journal of Production Research*, 53(1):220–237.

[Guardiola et al., 2009] Guardiola, L., Meca, A., and Puerto, J. (2009). Production-inventory games: A new class of totally balanced combinatorial optimization games. *Games and Economic Behavior*, 65(1):205–219.

[Gunal, 2012] Gunal, M. (2012). A guide for building hospital simulation models. *Health Systems*, 1(1):17–25.

[Guo et al., 2013] Guo, P., Leng, M., and Wang, Y. (2013). A fair staff allocation rule for the capacity pooling of multiple call centers. *Operations Research Letters*, 41(5):490–493.

[Harrison and López, 1999] Harrison, J. and López, M. (1999). Heavy traffic resource pooling in parallel-server systems. *Queueing Systems*, 33(4):339–368.

[Hartman and Dror, 1996] Hartman, B. and Dror, M. (1996). Cost allocation in continuous-review inventory models. *Naval Research Logistics*, 43(4):549–561.

[Hartman et al., 2000] Hartman, B., Dror, M., and Shaked, M. (2000). Cores of inventory centralization games. *Games and Economic Behavior*, 31(1):26–49.

[Iyer and Jain, 2004] Iyer, A. and Jain, A. (2004). Modeling the impact of merging capacity in production-inventory systems. *Management Science*, 50(8):1082–1094.

[Jouini, 2012] Jouini, O. (2012). Analysis of a last come first served queueing system with customer abandonment. *Computers & Operations Research*, 39:3040–3045.

[Jouini et al., 2008] Jouini, O., Dallery, Y., and Nait-Abdallah, R. (2008). Analysis of the impact of team-based organizations in call center management. *Management Science*, 54(2):400–414.

[Jouini et al., 2013] Jouini, O., Koole, G., and Roubos, A. (2013). Performance indicators for call centers with impatient customers. *IIE Transactions*, 45(3):341–354.

[Karsten et al., 2015a] Karsten, F., Slikker, M., and Borm, P. (2015a). Cost allocation rules for elastic single-attribute situations. Technical report, CentER Discussion Paper Series No. 2015-016.

[Karsten et al., 2011] Karsten, F., Slikker, M., and van Houtum, G. (2011). Analysis of resource pooling games via a new extension of the erlang loss function. Technical report, BETA Working Paper 344, Eindhoven University of Technology.

[Karsten et al., 2015b] Karsten, F., Slikker, M., and van Houtum, G. (2015b). Resource pooling and cost allocation among independent service providers. *Operation Research*, 63(2):476–488.

[Katta and Sethuraman, 2006] Katta, A. and Sethuraman, J. (2006). Cooperation in queues. Technical report, CORC technical reports, TR-2005-03, Columbia University, New York.

[Kayi and Ramaekers, 2010] Kayi, Ç. and Ramaekers, E. (2010). Characterizations of pareto-efficient, fair, and strategy-proof allocation rules in queueing problems. *Games and Economic Behavior*, 68(1):220–232.

[Kim and Kim, 2015] Kim, S. and Kim, S. (2015). Differentiated waiting time management according to patient class in an emergency care center using an open jackson network integrated with pooling and prioritizing. *Annals of Operations Research*, 230(1):35–55.

[Kleinrock, 1976] Kleinrock, L. (1976). *Queueing Systems Vol: II: Computer Applications*. John Wiley & Sons, Incorporated.

[Kostami and Ward, 2009] Kostami, V. and Ward, A. (2009). Managing service systems with an offline waiting option and customer abandonment. *Manufacturing & Service Operations Management*, 11(4):644–656.

[Little, 1961] Little, J. (1961). A proof for the queuing formula: L=$\lambda$w. *Operations research*, (3):383–387.

[Mandelbaum and Reiman, 1998] Mandelbaum, A. and Reiman, M. (1998). On pooling in queueing network. *Mangement Science*, 44(7):971–981.

[Mandelbaum and Zeltyn, 2009] Mandelbaum, A. and Zeltyn, S. (2009). Staffing many-server queues with impatient customers: constraint satisfaction in call centers. *Operations Research*, 57(5):1189–1205.

[Meca et al., 2004] Meca, A., Timmer, J., Garcı, I., Borm, P., et al. (2004). Inventory games. *European Journal of Operational Research*, 156(1):127–139.

[Mishra and Rangarajan, 2007] Mishra, D. and Rangarajan, B. (2007). Cost sharing in a job scheduling problem. *Social Choice and Welfare*, 29(3):369–382.

[Monahan, 2000] Monahan, G. (2000). *Management Decision Making: Spreadsheet Modeling, Analysis, and Application*, volume 1. Cambridge University Press.

[Müller et al., 2002] Müller, A., Scarsini, M., and Shaked, M. (2002). The newsvendor game has a nonempty core. *Games and Economic Behavior*, 38(1):118–126.

[Nagarajan and Sošić, 2008] Nagarajan, M. and Sošić, G. (2008). Game-theoretic analysis of cooperation among supply chain agents: Review and extensions. *European Journal of Operational Research*, 187(3):719–745.

[Özen et al., 2008] Özen, U., Fransoo, J., Norde, H., and Slikker, M. (2008). Cooperation between multiple newsvendors with warehouses. *Manufacturing & Service Operations Management*, 10(2):311–324.

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

[Peng et al., a] Peng, J., Jouini, O., and Jemai, Z. Cooperation in service systems with general service times. Under review.

[Peng et al., b] Peng, J., Jouini, O., and Jemai, Z. Service collaboration with impatient customers. Submitted.

[Peng et al., 2015] Peng, J., Jouini, O., Jemai, Z., and Dallery, Y. (2015). Service capacity pooling in M/G/1 service systems. In *Proceeding of the International Conference on Industrial Engineering and Systems Management (IEEE-IESM'2015), Seville, Spain*, pages 1097–1104.

[Pollaczeck, 1930] Pollaczeck, F. (1930). Uber eine aufgabe der wahrscheinlichkeitstheorie. *Mathematische Zeitschrift*, 32(1):64–100.

[Satir et al., 2012] Satir, B., Savasaneril, S., and Serin, Y. (2012). Pooling through lateral transshipments in service parts systems. *European journal of operational research*, 220(2):370–377.

[Schmeidler, 1969] Schmeidler, D. (1969). The nucleolus of a characteristic function game. *SIAM Journal on Applied Mathematics*, 17(6):1163–1170.

[Shaked and Shanthikumar, 1988] Shaked, M. and Shanthikumar, J. (1988). Stochastic convexity and its application. *Advances in Applied Probability*, 20:427–446.

[Shapley, 1952] Shapley, L. (1952). A value for n-person games. Technical report, DTIC Document.

[Shapley, 1971] Shapley, L. (1971). Cores of convex games. *International Journal of Game Theory*, 1(1):11–26.

[Sherali and Lunday, 2011] Sherali, H. and Lunday, B. (2011). Equitable apportionment of railcars within a pooling agreement for shipping automobiles. *Transportation Research Part E: Logistics and Transportation Review*, 47(2):263–283.

[Shi and Lian, 2016] Shi, Y. and Lian, Z. (2016). Optimization and strategic behavior in a passenger–taxi service system. *European Journal of Operational Research*, 249(3):1024–1032.

[Smith and Whitt, 1981] Smith, D. and Whitt, W. (1981). Resource sharing for efficiency in traffic systems. *Bell System Technical Journal*, 60(1):39–55.

[Stidham, 1970] Stidham, J. (1970). On the optimality of single-server queuing systems. *Operations Research*, 18(4):708–732.

[Tekin et al., 2014] Tekin, E., Hopp, W., and van Oyen, M. (2014). Pooling strategies for call center agent cross-training. *IIE Transactions*, 41(6):546–561.

[Thompson and Kwortnik, 2008] Thompson, G. and Kwortnik, R. (2008). Pooling restaurant reservations to increase service efficiency. *Journal of Service Research*, 10(4):335–346.

[Tijs, 1981] Tijs, S. (1981). Bounds for the core of a game and the t-value. *Game Theory and Mathematical Economics*, pages 123–132.

[Timmer et al., 2013] Timmer, J., Chessa, M., and Boucherie, R. (2013). Cooperation and game-theoretic cost allocation in stochastic inventory models with continuous review. *European Journal of Operational Research*, 231(3):567–576.

[Toivonen, 2014] Toivonen, M. (2014). Special section: new perspectives on sustainability. introduction. *ECONOMICS AND POLICY OF ENERGY AND THE ENVIRONMENT*, pages 19–27.

[USTR, 2014] USTR (2014). Service. https://ustr.gov/issue-areas/services-investment/services, Service and Investment.

[Van den Heuvel et al., 2007] Van den Heuvel, W., Borm, P., and Hamers, H. (2007). Economic lot-sizing games. *European Journal of Operational Research*, 176(2):1117–1130.

[Vanberkel et al., 2012] Vanberkel, P.T., B. R., Hans, E., et al. (2012). Efficiency evaluation for pooling resources in health care. *OR spectrum*, 34(2):371–390.

[Wallace and Whitt, 2004] Wallace, R. and Whitt, W. (2004). Resource pooling and staffing in call centers with skill-based routing. *Operations Research*, 7(4):276–294.

[Wallace and Whitt, 2005] Wallace, R. and Whitt, W. (2005). A staffing algorithm for call centers with skill-based routing. *Manufacturing & Service Operations Management*, 7(4):276–294.

[Whitt, 2006] Whitt, W. (2006). Fluid models for multiserver queues with abandonments. *Operations Research*, 54(1):37–54.

[Yu et al., 2015] Yu, Y., Benjaafar, S., and Gerchak, Y. (2015). Capacity sharing and cost allocation among independent firms with congestion. *Production and Operations Management*, (8):1285–1310.

[Yves, 1990] Yves, S. (1990). Population monotonic allocation schemes for cooperative games with transferable utility. *Game and Economic Behaviors*, 2(4):378–394.

[Zhang, 2009] Zhang, J. (2009). Cost allocation for joint replenishment models. *Operations Research*, 57(1):146–156.

**Titre :** Modèles de files d'attente pour l'analyse des stratégies de collaboration dans les systèmes de services

**Mots clés :** systèmes de service, mangement des opérations, pooling des ressources, files d'attente, jeux coopératifs, simulation

**Résumé :** Au cours des vingt dernières années, le secteur des services est devenu le secteur le plus important en nombre d'actifs occupés dans l'économie mondiale, en particulier dans les pays développés. Par ailleurs, la concurrence et la coopération dans le secteur des services sont devenues de plus en plus populaires dans le contexte de la mondialisation économique. Comment collaborer avec un accord gagnant-gagnant apporte une source fertile de problèmes de management des opérations dans le domaine des services. Dans cette thèse, nous étudions des stratégies de collaboration dans des systèmes de services homogènes. Nous nous concentrons en particulier sur les stratégies de pooling des ressources de service.
Dans les deux premières parties, nous étudions le problème de partage des coûts entre les fournisseurs de services indépendants avec des temps de service qui suivent une distribution générale et en tenant compte de l'abandon des clients. Nous modélisons à la fois chaque fournisseur et la coalition coopérative comme des files d'attente avec serveur unique, et spécialisons les stratégies de pooling avec les capacités de service fixes et modifiables.
Dans la dernière partie, nous abordons le problème de pooling dans le cadre multiserveur pour évaluer la qualité de l'hypothèse 'super-serveur'. Nous étudions numériquement l'impact de la variabilité de la durée de service et l'abandon des clients sur les jeux de mise en commun des ressources. Nous comparons aussi les partages des coûts entre le système de 'super-serveur' et multiserveur.

**Title :** Queueing approaches for the analysis of collaboration strategies in service systems

**Keywords :** service systems, operations management, resource pooling, queueing theory, cooperative game theory, simulation

**Abstract:** In past twenty years, the service sector has emerged as the primary sector in the world economy, especially in developed countries. Competition and cooperation in service industries have become more and more popular in the context of economic globalization. How to operate the collaboration with a win-win agreement brings a fertile source of operations management issues in service science. In this thesis, we study collaborations between homogeneous service systems in terms of resource pooling strategies.
In the first two parts, we investigate the cost-sharing problem among independent service providers with general service times and accounting for the customer abandonment. We model both the service provider and the cooperative coalition as single server queues, and specialize the capacity pooling strategies with the fixed and optimized service capacities.
Finally, we address the service pooling problem in the multi-server pooling setting to assess the quality of the 'super-server' assumption. We numerically investigate the impact of service duration variability and customer abandonment on the pooling game. We compare between cost-sharing results of the two resource pooling concepts, with or without the 'super-server' assumptions.