Capacity Sharing and Cost Allocation among Independent Firms with Congestion

Yimin Yu
Department of Management Sciences
City University of Hong Kong
Kowloon, Hong Kong
yiminyu@cityu.edu.hk

Saif Benjaafar
Department of Industrial and Systems Engineering
University of Minnesota Minneapolis, MN 55455
saif@umn.edu

Yigal Gerchak
Department of Industrial Engineering
Tel Aviv University, Tel Aviv, Israel
ygerchak@eng.tau.ac.il

Abstract

We analyze the benefit of production/service capacity sharing for a set of independent firms. Firms have the choice of either operating their own production/service facilities or investing in a facility that is shared. Facilities are modeled as queueing systems with finite service rates. Firms decide on capacity levels (the service rate) to minimize delay costs and capacity investment costs possibly subject to service level constraints on delay. If firms decide to operate a shared facility they must also decide on a scheme for sharing the capacity cost. We formulate the problem as a cooperative game and identify settings under which capacity sharing is beneficial and there is a cost allocation that is in the core under either the FCFS policy or an optimal priority policy. We show that capacity sharing may not be beneficial in settings where firms have heterogeneous work contents and service variabilities. In such cases, we specify conditions under which capacity sharing may still be beneficial for a subset of the firms.

Key words: Capacity sharing, queueing systems, joint ventures, cost allocation, cooperative game theory

1 Introduction

Capacity sharing refers to the fulfillment of demand that arises from multiple sources from a single facility instead of facilities dedicated to each demand source. In a system without capacity sharing, each dedicated facility fulfills its own demand relying solely on its capacity. It has long been known that capacity sharing can be beneficial when demand is random. This benefit can be in the form of improved service quality with the same amount of capacity or in the form of less capacity needed to provide the same quality of service. Capacity sharing can also be beneficial when there are economies of scale associated with acquiring capacity or fulfilling demand. These benefits have been shown to be true for various forms of capacity, including manufacturing, service, and inventory.

Capacity sharing has been studied mostly in situations where a single firm, or a sub-division within a firm, owns all the capacity in the system, and has responsibility for serving all the demand. This firm makes the decision about whether or not to share capacity and how much capacity to acquire. In this paper, we consider a system with n independent firms, or sub-divisions within a firm, each facing its own demand and each having the option of either operating its own independent facility or joining some or all the other firms in a shared facility. The firms may vary in their demand levels and in their tolerance for capacity shortage. If some or all of the firms decide to share capacity, they must also decide on how to allocate the cost of the shared facility. They must do so in a manner that benefits everyone and prevents any of the firms from defecting and perhaps sharing a facility with a subset of the firms or staying on their own. Hence, firms that contribute more to the cost of the shared facility (because of their higher usage of capacity or lower tolerance for capacity shortage) are expected to pay a greater share of total cost.

In this paper, we consider applications where facilities can be modeled as queueing systems. Demand for each firm consists of an independent stream of customers (or orders) that arrive continuously over time with random inter-arrival times. Customers are processed at each facility one at a time with stochastic service times. The capacity at each facility is determined by the rate at which customers can be processed. Because customers are processed one at a time and because customer arrivals and processing times are random, congestion arises and customers can experience delay prior to processing (if a customer arrives and finds the service facility busy, the customer must wait for service). Each firm can install and operate its own facility where its customers are processed. Firms make decisions about how much service capacity to acquire in order to minimize two types of costs, delay cost due to customers spending time at the facility prior to completing service and capacity investment cost, subject to a constraint on the amount of delay or waiting time

that customers experience. Alternatively, firms may choose to collectively operate a shared facility. In that case, in addition to determining the optimal amount of capacity (taking into account the delay costs and service levels of all the firms), the firms must also determine how the corresponding costs must be allocated. In both cases, of either shared or individual facilities, the facilities are modeled as single server queues, with capacity determined by the associated service rate.

Capacity sharing among independent firms in the presence of congestion, and with heterogeneous delay costs and service level requirements, arises in a variety of settings. For example, firms (or sub-divisions within a firm) can decide to share support services, such as repair and maintenance facilities, instead of investing in facilities of their own [1]. In the case of repairs, repair requests for each firm arise randomly over time (because of the randomness in breakdowns) with repair times that can be stochastic. Given the limited repair capacity, this can lead to congestion. Each firm may have its own delay cost (e.g., costs corresponding to the opportunity cost of equipment down time) or may have a specified service level it guarantees to its own customers. Capacity in this case is determined by the speed with which repairs can be undertaken (e.g., the processing rate of the main bottleneck process). Other examples include common services, such as printing, testing, and rapid prototyping, which sub-divisions within the same firm may choose to share. In this case too, the variability in the arrival of service requests and processing times can lead to congestion. Capacity, in terms of the speed with which service requests can be processed, must again take into account the requirements of the different subdivisions, including their individual sensitivity to delay.

The main contributions of our paper are summarized below.

- We provide a framework for modeling capacity sharing in queueing systems with independent firms. To our knowledge, our paper is among the first to model the issue of cooperation and capacity sharing in a queueing context.

- Under the M/M/1 setting with the FCFS policy, we formulate capacity sharing as a cooperative game, in which the participating firms optimize shared capacity taking into account the characteristics of individual firms, their delay costs and their service level requirements. We characterize settings where the *core* of the game is non-empty. That is, there exists a capacity cost allocation rule for which all the firms are better off than under any other alternative sharing arrangement, including being on their own.

[1] For example, several airlines share repair and maintenance facilities; see recent announcements by Air France and Lufthansa http://www.lufthansa-technik.com/spairliners and Delta and Aero Mexico http://news.delta.com/index.php?s=43&item=1698.

- When the core exists, we identify a simple and easy-to-implement allocation rule with desirable properties that is in the core. The allocation rule charges every firm the cost of capacity for which it is directly responsible, its own delay cost, and a fraction of buffer capacity cost that is consistent with its contribution to this cost.

- We consider systems that operate under an optimal priority policy. We show that there exists a capacity cost sharing scheme that is in core. We accomplish this by showing that the corresponding cooperative game is submodular.

- Under the FCFS policy, we also characterize settings where the core may not exist because capacity sharing may not be beneficial. These settings include systems where service levels are specified in terms of waiting time instead of total delay and cases where the firms are heterogeneous in their characteristics, including their work contents, service time variability, and delay costs. All these indicate that capacity sharing should be considered with caution in contrast to the common belief that risk pooling is beneficial. For these cases, we characterize conditions, under which capacity sharing may still be beneficial for a subset of the firms. These conditions provide insights into the characteristics of firms that would benefit from forming sub-coalitions.

- We extend our results to a variety of queueing systems, including M/G/1 queues, and GI/G/1 queues. In doing so, we extend known results regarding the benefit of capacity pooling in each system by endogenizing capacity instead of assuming it remains constant with and without capacity sharing.

We should note that there is a rich literature that models manufacturing and service systems as queueing systems (see Sections 2 and 3 for further discussion). Surprisingly very little of this literature addresses the issue of cooperation and capacity sharing when there are independent firms. Therefore, we view our paper as a step toward a more comprehensive examination of the issue of cooperation in queueing systems, whether it arises in manufacturing, services or elsewhere. We also view it as a contribution, in the form of a potentially rich application domain, to the literature on cooperative games.

The rest of the paper is organized as follows. In Section 2, we provide a brief review of related literature. In Section 3, we treat the case with no capacity sharing. In Section 4, we analyze the case with capacity sharing. In Section 5, we consider capacity sharing when there are service priorities. In Section 6, we extend our analysis to systems with heterogeneous work contents and systems with general arrival and service processes. In Section 7, we offer concluding comments.

2 Related Literature

There is a rich literature on capacity *pooling* in queueing systems, with applications ranging from manufacturing and service operations to telecommunications systems to computer networks. This literature can be classified broadly as relating to either the *pooling of service rates* or the *pooling of servers*. Server rate pooling refers to the consolidation of multiple servers into a single one with a faster rate (e.g., N servers, each with service rate μ and demand rate λ, are replaced by a single server with service rate $N\mu$ and demand rate $N\lambda$). Server pooling on the other hand refers to placing multiple servers in a single facility from which all demand streams are served (e.g., N single server queues are replaced by a single multi-server queue with N servers and a demand rate $N\lambda$).

Kleinrock (1976) discusses various examples of both types of pooling. Stidham (1970) considers a design problem where the decision variables are the number of parallel servers and the service rate of each server. Smith and Whitt (1981) and Benjaafar (1995) show that server pooling, when the number of servers is exogenously determined, is beneficial as long as all customers have identical service time distributions. Buzacott (1996) considers the pooling of N servers in series, with each server dedicated to one task, into N parallel servers, with each server carrying out all the tasks. Mandelbaum and Reiman (1998) consider the pooling of general Jackson networks into single server queues with phase-type service time distributions.

Tekin et al. (2009) use approximations to evaluate the benefit of partitioning servers in multiple pools instead of a single large one. Sheikhzadeh et al. (1998), Gurumurthi and Benjaafar (2004) and Jordan et al. (2005) study the *chaining* of servers, where each server can process customers from two customer streams and each customer can be routed to two servers. They show that in systems with homogeneous demand rates and service time requirements, chaining can achieve most of the benefits of total server pooling; see also Hopp et al. (2004), Iravani et al. (2004), Bassambo et al. (2008), Aksin et al. (2008), Wallace and Whitt (2005) and the references therein. These papers belong to the growing literature on queueing systems with server flexibility (or cross-training); see Jouini et al. (2008), Aksin et al. (2005) and Koole and Pot (2005) for recent reviews.

The treatment in this paper is different from the above literature in three important aspects. First, we do not assume that there is a single decision maker that determines whether or not to pool. Instead, we consider multiple firms that decide independently on either operating their own facilities or sharing one with other firms (pooling here does not imply a merger however). Second, we do not assume that service capacity is exogenously given. We allow for this to be an outcome of an optimization carried out by the firms either individually or jointly. Third, we are concerned

with identifying cost allocation schemes under which all firms prefer a single shared facility to any other capacity sharing arrangement, including remaining on their own.

The literature dealing with capacity sharing in the context of independent firms is limited. Gonzalez and Herrero (2004), and also Garcia-Sanz et al. (2008), consider a special case of the M/M/1 model we consider. However in both cases, they do not optimize capacity (before or after pooling) and do not consider the delay cost. In our case, the presence of delay costs significantly complicates the process of cost allocation since we seek allocations that could allow for each firm to absorb its own cost of delay. Anily and Haviv (2010) treat a related M/M/1 model where the issue is how to allocate delay cost to ensure that the allocation is in the core. They show that a Shapley allocation based on the service levels is in the core. In this literature, the common approach is to use concavity as a basis for proving that the core exists and that a Shapley allocation is in the core; see Gonzalez and Herrero (2004), Garcia-Sanz et al. (2008), and Anily and Haviv (2010). In contrast to the above literature, a Shapley allocation may not be in the core in our case. This is in part due to our treatment of capacity as endogenous and to the requirement that each firm absorbs its own delay cost and only capacity costs are allocated among the firms. We also treat queueing systems other than the M/M/1 queue, including queues with service priorities, M/G/1 queues, and GI/G/1 queues.

Our work is of course related to the vast literature on cooperative game theory and, more broadly, the economics of coalition formation and joint ventures; see Moulin (1995) for a general introduction to the topic. Some of this literature has focused on cooperation involving sequencing and scheduling; see for example Moulin and Stong (2002), Maniquet (2003), and Katta and Sethuraman (2006). This literature sometimes refers to these problems as queueing problems. However, they typically involve a finite population of customers who simultaneously arrive to the system, and therefore are not concerned with steady state behavior and congestion in the way that we are in this paper. In Operations Management, there is growing literature that applies cooperative game theory to joint ordering problems, particularly in the context of economic order quantity models (see Anily and Haviv (2007), Dror and Hartman (2007) and the many references therein), economic lot sizing models (see for example van den Heuvel (2007) and Chen and Zhang (2009), among others), and news-vendor models (see Muller et al. (2002), Nagarajan and Sošić (2008), Kemahlioglu-Ziya (2004), Chen and Zhang (2007), and Hanany and Gerchak (2008) and the references therein).

Finally, we should note that there is a rich literature on outsourcing where multiple firms may be served by the same supplier, including for settings where the outsourcing supplier is modeled as a queueing system; see for example, Cachon and Harker (2002), Allon and Federgruen (2006), Gans

and Zhou (2007), and Benjaafar et al. (2007). In general, the focus of this literature is different as it does not deal with cost allocation or coalition formation.

3 Systems without Capacity Sharing

Consider a system consisting of a set $\mathcal{N} = \{1, ..., n\}$ of n firms. Firm i, $i \in \mathcal{N}$, faces an independent demand stream with customers arriving according to a Poisson process with rate λ_i (we treat more general arrival processes in Section 6). When firms operate independently, each firm invests in a separate service facility and chooses a certain level of capacity in the form of a service rate. We refer to this scenario as the scenario without capacity sharing. Once the facilities are built, each firm serves its customers from its own facility one at a time on a first-come, first-served (FCFS) basis. We assume service times are independent and identically distributed random variables denoted by X_i where X_i is of the form Y/μ_i and Y is a random variable that is exponentially distributed with a mean equal to 1. Hence, service time is also exponentially distributed with mean $E[X_i] = 1/\mu_i$. The parameter μ_i, $(\mu_i > 0)$ is a scaling parameter that corresponds to the service rate or capacity.

The random variable Y can be viewed as the work content associated with each customer. We assume that the work content is homogeneous across firms. This assumption is justified if firms provide service to similar customers (e.g., repairing similar equipments in the case of a maintenance facility). Given the exponential nature of both customer inter-arrival times and service times, each firm behaves like an M/M/1 queue. There is a significant literature on the economics of queues in *competitive* settings that primarily focuses on the M/M/1 queue (and where the service rate is the decision variable); see Hassin and Haviv (2003) for a review of that literature and see Cachon and Harker (2002), Cachon and Zhang (2007), Benjaafar et al. (2007), and Allon and Federgruen (2007), among many others, for example applications. Our treatment of the M/M/1 queue is consistent with assumptions made in that literature and can be viewed as complementing it for cooperative settings.

We assume that service rate can be varied continuously and that firms incur a capacity cost c per unit of service rate per unit time. This is justified in settings where capacity can be continuously scaled over a sufficiently large interval. It is consistent with treatments elsewhere in the literature (see for example Kalai et al. (1992), Mendelson and Whang (1990), Ha (2001), Allon and Federgruen (2007, 2008), Cachon and Zhang (2007), and the vast literature reviewed therein). This assumption can also be found in the significant literature on capacity planning, as noted recently by Bassambo et al. (2008). The assumption of linear capacity cost implies that there are neither economies nor

diseconomies of scale. This is an important case that has been widely studied in the literature (see Allon and Federgruen (2007, 2008), Dewan and Mendelson (1990), Stidham (1992), Cachon and Harker (2002), and Bassambo et al. (2008) among others), leads to tractable analysis, and provides a useful benchmark for other cost structures.

We assume that the demand rate for each firm is known. This of course does not mean that demand is deterministic. Inter-arrival times between consecutive customers are stochastic. Therefore, the number of customers that arrive over a given period of time is random. The assumption of known demand rate is consistent with most of the existing literature on capacity planning in queueing systems (and indeed in most of the queueing literature); see for example Kleinrock (1976), Cachon and Harker (2002), Bassambo et al. (2008), and Allon and Federgruen (2007, 2008), among many others.

The objective of each firm is to minimize its capacity investment while limiting the amount of delay, as measured by either total time in system, or waiting time in the queue, its customers experience. Limiting customer delay can be achieved by enforcing a service level constraint or by associating a cost with the amount of delay customers experience. A service level constraint may take several forms, including a constraint on the probability of customer delay not exceeding a specified threshold, or a constraint on expected delay not exceeding a certain maximum. Service level constraints are managerial decisions that typically reflect either a position in the marketplace that a firm would like to take or contractual obligations that a firm has negotiated with its customers. We assume that all firms choose the same type of service level constraints since they are in the same industry.

Delay costs can reflect either direct or indirect costs. Direct costs are penalties incurred by the firm due to delays experienced by its customers (for example, payments to customers to compensate for the total time that they spend in the system) or indirect costs due to loss of customer goodwill. Hence, delay costs are not unlike backorder costs, common in inventory settings (Zipkin 2000). The use of delay costs and service levels are both common in the literature; see for example Dewan and Mendelson (1990), Mendelson and Whang (1990), Ha (1998, 2001), Allon and Federgruen (2007, 2008) and the references therein.

In this paper, we consider the case where a unit delay cost h_i is incurred for each unit of time a customer spends in the system (time either in the queue or in service in steady state) and the objective is to minimize the long run expected delay cost. Moreover, we consider the case where service level is expressed in terms of a probability that delay in the system for each customer, which we define as the sum of waiting time in the queue and time in service, does not exceed a

specified threshold. This measure is consistent with service levels used elsewhere in the literature; see for example Allon and Federgruen (2007, 2008), among others. We also consider service levels expressed in terms of waiting time the queue alone. Service level measured in terms of total time in system is appropriate in applications such as computing, telecommunication, and manufacturing where customers are concerned about the total fulfillment of their orders/service requests. Service level measured in terms of waiting time is appropriate in settings, such as call centers and other service systems, where customers are particularly sensitive to time spent in the queue.

Let $z_i(\mu_i)$ denote the expected total cost incurred by firm i given a service rate μ_i (for stability, we assume that $\lambda_i/\mu_i < 1$). Let W_i, a random variable, denote the delay (waiting time in the queue + service time) that a customer of firm i experiences and $P(W_i \leq w_0)$ the probability that customer delay does not exceed w_0 where $w_0 \geq 0$ (we will consider later the case where service level is expressed in terms of waiting time). The problem faced by firm i can then be stated as follows

$$\text{Minimize} \quad z_i(\mu_i) = c\mu_i + \frac{h_i\lambda_i}{\mu_i - \lambda_i} \tag{1}$$

$$\text{subject to } P(W_i \leq w_0) = 1 - e^{-(\mu_i - \lambda_i)w_0} \geq \alpha_i, \tag{2}$$

$$\lambda_i/\mu_i \leq 1.$$

The objective function in the above optimization problem consists of two terms: a capacity cost term and a delay cost term, where the decision variable is the capacity level of firm i as determined by the service rate μ_i. The formulation captures two important special cases: (1) the case where $\alpha_i = 0$ for all $i \in \mathcal{N}$ and (2) the case where $h_i = 0$ for all $i \in \mathcal{N}$. The first corresponds to a pure cost-based formulation with no constraints on service levels, while the second corresponds to a service level-based formulation with no delay costs. In the absence of service level constraints, the optimal capacity level μ_i^* can be obtained from the first order condition of optimality, since z_i is convex in μ_i, as

$$\mu_i^* = \lambda_i + \sqrt{\frac{h_i\lambda_i}{c}}. \tag{3}$$

In systems with service level constraints but no delay costs, the optimal capacity level is given by the smallest μ_i that satisfies inequality (2). This leads to the following optimal capacity level

$$\mu_i^* = \lambda_i + \frac{\ln(\frac{1}{1-\alpha_i})}{w_0}. \tag{4}$$

Surprisingly, the buffer capacity $\frac{\ln(\frac{1}{1-\alpha_i})}{w_0}$ is independent of the demand rate, a result of the fact that the delay distribution depends on λ_i and μ_i through $\mu_i - \lambda_i$ only. This feature is also present if service levels are specified in terms of expected delay. If we let α_i now denote the threshold on the maximum expected delay, then the service level constraint is given by $\frac{1}{\mu_i - \lambda_i} \leq \alpha_i$. This leads to an optimal capacity given by $\mu_i^* = \lambda_i + \frac{1}{\alpha_i}$. Note that the buffer capacity is again independent of the demand rate.

In both cases, the optimal capacity is the sum of two components. The first corresponds to the demand rate, λ_i (since all demand must be satisfied) while the second corresponds to *buffer* capacity that increases in either the ratio $\frac{h_i\lambda_i}{c}$ or the service level α_i. The expressions in equations (3) and (4) are not new. Similar expressions have been derived elsewhere; see for example Kleinrock (1976), Allon and Federegruen (2008) and Hassin and Haviv (2003).

In the general case, with both delay costs and service level constraints, the optimal capacity level is given by

$$\mu_i^* = \lambda_i + \eta_i, \tag{5}$$

where

$$\eta_i = \max\{\frac{\ln(\frac{1}{1-\alpha_i})}{w_0}, \sqrt{\frac{h_i\lambda_i}{c}}\}. \tag{6}$$

Substituting μ_i^* in (1), we obtain the optimal expected cost for firm i as

$$z_i^* = c(\lambda_i + \eta_i) + \frac{h_i\lambda_i}{\eta_i}.$$

This leads to a total system cost of $z_{1,\ldots,n}^* = \sum_{i\in\mathcal{N}} z_i^*$. In systems where $\sqrt{\frac{h_i\lambda_i}{c}} \geq \frac{\ln(\frac{1}{1-\alpha_i})}{w_0}$ for all $i \in \mathcal{N}$, the optimal cost simplifies to $z_i^* = c\lambda_i + 2\sqrt{h_i\lambda_i c}$. This leads to a total system cost, $z_{1,\ldots,n}^*$, given by $z_{1,\ldots,n}^* = c\sum_{i\in\mathcal{N}}\lambda_i + 2\sum_{i\in\mathcal{N}}\sqrt{h_i\lambda_i c}$. In the case of identical firms, with $\lambda_i = \lambda$ and $h_i = h$ for all $i \in \mathcal{N}$, the optimal total cost reduces to $z_{1,\ldots,n}^* = cn\lambda + 2n\sqrt{h\lambda c}$, and the total capacity in the system to $\sum_{i\in\mathcal{N}}\mu_i^* = n(\lambda + \sqrt{\frac{h\lambda}{c}})$. As we can see, both the optimal cost and the optimal buffer capacity in the system increase linearly in the number of firms n. Similar observations can be made for systems in which $\sqrt{\frac{h\lambda}{c}} \leq \frac{\ln(\frac{1}{1-\alpha})}{w_0}$. That is, in this case too, both the optimal cost and the optimal buffer capacity in the system increase linearly in n when the firms have identical cost, service level, and demand parameters.

Next we consider the case where the service level is specified in terms of waiting time in the queue. Let Q_i, a random variable, denote the time a customer of firm i spends waiting in the queue before service starts and let $P(Q_i \leq q_0)$ be the probability that customer waiting time does not

exceed q_0 where $q_0 \geq 0$. The problem faced by firm i can then be restated as

$$\text{Minimize} \quad z_i(\mu_i) = c\mu_i + \frac{h_i\lambda_i}{\mu_i - \lambda_i} \tag{7}$$

$$\text{subject to} \quad P(Q_i \leq q_0) = 1 - \frac{\lambda_i}{\mu_i} e^{-(\mu_i - \lambda_i)q_0} \geq \alpha_i, \tag{8}$$

$$\lambda_i/\mu_i \leq 1.$$

In the absence of service level constraints, the optimal capacity level is given by (3). In systems with service level constraints but no delay costs, the optimal capacity level is given by the smallest μ_i that satisfies inequality (8) or, equivalently, the optimal capacity is the solution to the following equation

$$\ln(\lambda_i q_0) + \lambda_i q_0 - \ln(1 - \alpha_i) = \ln(\mu_i q_0) + \mu_i q_0. \tag{9}$$

Unfortunately, there is no explicit solution for the above equation. However, we are able to show the following important result.

Lemma 3.1 *Let $\bar{\mu}_i(\alpha_i, \lambda_i)$ be the solution to (9) and $\bar{\eta}_i(\alpha_i, \lambda_i) = \bar{\mu}_i(\alpha_i, \lambda_i) - \lambda_i$, the amount of buffer capacity ($\bar{\mu}_i(\alpha_i, \lambda_i) > \lambda_i$). Then, given α_i, $\bar{\eta}_i(\alpha_i, \lambda_i)$ is nondecreasing in λ_i, with $\bar{\eta}_i(\alpha_i, \lambda_i)/\lambda_i$ being nonincreasing in λ_i.*

(The proof of this and of all subsequent results can be found in the Appendix A).

This lemma indicates that, in contrast to the case where service level is delay-based, buffer capacity is increasing in demand, although the rate of increase is less than one. This result is due to the fact that the delay is exponentially distributed with the rate of the buffer capacity while the waiting is not exponentially distributed. As we will see in the next section, this significantly affects the benefit derived from capacity sharing.

4 Systems with Capacity Sharing

In this section, we consider the scenario where the firms decide to form a coalition and invest in a single shared facility (a joint venture) from which the demand of all the firms will then be satisfied. We assume that the rules governing the joint venture (as negotiated by members of the coalition) require that the choice of capacity, in the form of a service rate, for the shared facility takes into account the demand levels of each member of the coalition, their delay costs, and their service level requirements. In particular, we assume that the service rate is chosen by the managers of the

joint venture so that it minimizes the total cost for the coalition (the sum of expected delay costs experienced by customers of all the firms and the cost of capacity) and satisfies all service level constraints. We assume that all members of the coalition are truthful in their reporting of their demand rates, delay costs, and service levels. We assume throughout that, although independent, the firms are not competitors so that their demands are exogenously determined and are not affected by decisions made by any of the firms.

The assumption of full information applies to settings where the information is public and can be independently verified by all the firms. For example, delay penalties and service level guarantees could be publicly advertised by the firms themselves as part of their marketing strategy. In some cases, delay penalties and service levels may also adhere to well-known industry standards. In settings where delay costs are directly incurred by the shared facility (e.g., the shared facility is responsible for handling delay penalty payments to the customers), firms would also need to provide the shared facility with the correct delay costs. Similarly, service levels must be known to the shared facility if contractual agreements with the customers regarding service levels are handled directly by the shared facility. Demand rates are in most cases verifiable since demand would eventually be satisfied from the shared facility. Firms can be induced to disclose their true demand rates by imposing high penalties if the originally reported rates are higher than the realized rates (measured over a sufficiently long period of time) once the facility is in operation. The assumption of full information is of course applicable to the case where the firms are all subdivisions of a single large firm.

We refer to the service rate in the shared facility from which the demand of all firms is satisfied as $\mu_{\mathcal{N}}$ (from heretofore, we shall index parameters associated with a set of firms with the name of that set while parameters associated with individual firms with the name of the firm). Because the superposition of independent Poisson processes is also a Poisson process, the demand process at the shared facility is Poisson with rate $\sum_{i\in\mathcal{N}} \lambda_i$. Similarly, because the work content for each customer regardless of its firm is exponentially distributed, the processing time at the shared facility is a random variable $X_{\mathcal{N}} = Y/\mu_{\mathcal{N}}$ with the exponential distribution and mean $1/\mu_{\mathcal{N}}$. We assume that customers regardless of their firm affiliation are served in a FCFS fashion. Hence, the system with the shared facility behaves again as an M/M/1 queue.

4.1 Capacity Optimization

First, we consider the case where the service level is specified in terms of delay (see Section 4.3 for a discussion of waiting time-based service levels). We assume that the terms of the joint venture

between the participating firms in the coalition require that the shared facility invests in capacity so as to minimize the total cost to the coalition while satisfying the service level constraint of each firm. The total cost to the coalition consists of the sum of capacity cost and expected delay cost (experienced by customers of all the firms over the long run). Satisfying the service level constraints of all the firms requires satisfying the highest of these service level constraints. If we let $z_{\mathcal{N}}(\mu_{\mathcal{N}})$ denote total system cost and let $W_{\mathcal{N}}$, a random variable, refer to customer delay, then the capacity optimization problem can be stated as follows:

$$\begin{aligned} \text{Minimize } z_{\mathcal{N}}(\mu_{\mathcal{N}}) =& c\mu_{\mathcal{N}} + \frac{\sum_{i\in\mathcal{N}} h_i\lambda_i}{\mu_{\mathcal{N}} - \sum_{i\in\mathcal{N}} \lambda_i} \\ \text{subject to } & P(W_{\mathcal{N}} \le w_0) = 1 - e^{-(\mu_{\mathcal{N}} - \sum_{i\in\mathcal{N}} \lambda_i)w_0} \ge \alpha_{\mathcal{N}}, \\ & \sum_{i\in\mathcal{N}} \lambda_i/\mu_{\mathcal{N}} \le 1, \end{aligned} \tag{10}$$

where $\alpha_{\mathcal{N}} = \max(\alpha_1, ..., \alpha_n)$. Then, the optimal capacity is given by

$$\mu^*_{\mathcal{N}} = \sum_{i\in\mathcal{N}} \lambda_i + \eta_{\mathcal{N}},$$

where

$$\eta_{\mathcal{N}} = \max\{\frac{\ln(\frac{1}{1-\alpha_{\mathcal{N}}})}{w_0}, \sqrt{\frac{\sum_{i\in\mathcal{N}} h_i\lambda_i}{c}}\}.$$

Similar to the system without capacity sharing, the optimal capacity consists of two components. The first corresponds to the total demand rate, while the second to buffer capacity which, in this case, increases in either the sum of the ratios $\frac{h_i\lambda_i}{c}$ or the maximum service level $\alpha_{\mathcal{N}}$. The results are similar if the service level is specified in terms of a threshold on expected delay. In that case the optimal buffer capacity is given by $\eta_{\mathcal{N}} = \max\{\frac{1}{\alpha_{\mathcal{N}}}, \sqrt{\frac{\sum_{i\in\mathcal{N}} h_i\lambda_i}{c}}\}$.

The following theorem shows that by investing in a shared facility, the firms are able to reduce total cost in the system while investing in less capacity.

Theorem 4.1 *$z^*_{\mathcal{N}} \le z^*_{1,...,n}$ and $\mu^*_{\mathcal{N}} \le \sum_{i=1}^n \mu^*_i$, where $z^*_{\mathcal{N}}$ is the optimal cost in the shared facility.*

The potential magnitude of the savings from capacity sharing can be more easily seen in a system with identical firms where $\alpha_i = \alpha$, $h_i = h$, and $\lambda_i = \lambda$ for all $i \in \mathcal{N}$. Consider the case where $\sqrt{\frac{h\lambda}{c}} \ge \frac{\ln(\frac{1}{1-\alpha})}{w_0}$. This leads to $\mu^*_{\mathcal{N}} = n\lambda + \sqrt{\frac{nh\lambda}{c}}$, $z^*_{\mathcal{N}} = cn\lambda + 2\sqrt{cnh\lambda}$, and $E(W^*_{\mathcal{N}}) = \sqrt{\frac{c}{nh\lambda}}$ from which we can observe that both buffer capacity and expected delay, and consequently delay cost, are reduced by a factor of a square root of n (relative to those observed in the case of no capacity sharing). In

the case where $\sqrt{\frac{nh\lambda}{c}} \leq \frac{\ln(\frac{1}{1-\alpha})}{w_0}$, we have $\mu^*_{\mathcal{N}} = n\lambda + \frac{\ln(\frac{1}{1-\alpha})}{w_0}$, $z^*_{\mathcal{N}} = c(n\lambda + \frac{\ln(\frac{1}{1-\alpha})}{w_0}) + \frac{nh\lambda w_0}{\ln(\frac{1}{1-\alpha})}$, and $E(W^*_{\mathcal{N}}) = \frac{w_0}{\ln(\frac{1}{1-\alpha})}$. Here, the magnitude of savings on capacity is even larger with buffer capacity reduced by a factor of n, but expected delay remains unchanged from the case without capacity sharing.

4.2 Cost Sharing

We have so far showed that capacity sharing is system-optimal. However, whether or not it is also optimal for individual firms depends on how the cost of the shared facility is allocated among the firms. We assume that each firm incurs its own delay cost and pays a fraction of capacity cost. A firm would prefer the shared facility if the sum of its share of capacity cost and its long run expected delay cost is lower than the cost it would incur without capacity sharing. Moreover, in many settings, the choice is not just between a single facility shared among all firms or facilities operated individually by each firm. There may instead be a range of facility sharing options. For example, a firm may find it more advantageous to share capacity with only a subset of the firms. This could lead firms to form groupings around multiple smaller shared facilities. A single shared facility would be preferred by all firms only if there exists a cost allocation under which the firms are better off than under any other capacity sharing arrangement, including operating individual facilities. Hence, it is desirable that the cost allocation for the shared would be designed so that it deters firms from breaking away and engaging in other facility sharing arrangements.

The problem of determining whether or not there exists a cost allocation scheme under which firms prefer to share a single facility to any other facility sharing configuration can be formulated as a *cooperative game* among the independent firms in the set $\mathcal{N}$. Consistent with standard terminology from cooperative game theory, let us refer to the subset of firms $\mathcal{J} \subseteq \mathcal{N}$ as *coalition* $\mathcal{J}$ and to the set $\mathcal{N}$, the largest coalition, as the *grand coalition.* A cooperative game is then defined by a characteristic function which specifies the value associated with each coalition $\mathcal{J}$. In our context, this corresponds to the total expected cost associated with a subset of firms $\mathcal{J}$ sharing a single facility. We refer to this cost as $z^*_{\mathcal{J}}$, where $z^*_{\mathcal{J}} \equiv z_{\mathcal{J}}(\mu^*_J)$. A vector $\phi = (\phi_1, \cdots, \phi_n)$ is called an allocation rule if ϕ_i corresponds to the portion of total expected cost in the grand coalition that is incurred by firm i. If $\sum_{i=1}^n \phi_i = z^*_{\mathcal{N}}$, then the allocation rule is said to be efficient. An allocation rule is said to be individually rational if $\phi_i \leq z^*_i$ and to be stable for a coalition J if $\sum_{i\in\mathcal{J}} \phi_i \leq z^*_{\mathcal{J}}$.

An allocation is said to be a member of the core if it satisfies the following inequalities:

$$\sum_{i\in\mathcal{J}} \phi_i \leq z^*_{\mathcal{J}},\ \forall \mathcal{J} \subseteq \mathcal{N}, \tag{11}$$

$$\sum_{i\in\mathcal{N}} \phi_i = z^*_{\mathcal{N}}. \tag{12}$$

When an allocation rule is in the core, no subset of players would want to secede from the grand coalition and form smaller coalitions, including being on their own. Hence the existence of an allocation rule that is in the core (the core is non-empty) is sufficient in our context to show that it is optimal for all the firms to share a single facility. This single facility is a superior arrangement to any other arrangement that may involve a set of partially pooled facilities shared among multiple subsets of the firms.

In addition to the requirement of being in the core, it is desirable for an allocation rule to be perceived as *fair*. In general, a fair allocation is one that assigns a higher portion of total cost to firms whose membership in the coalition contribute more to total cost. In particular, everything else being equal, firms with higher demand rates, higher delay costs, or higher service levels should pay a greater portion of total cost. In what follows, we show that a relatively simple allocation rule has both the properties of being in the core and satisfying the above intuitive notions about fairness (for a more extensive discussion of fairness in cost allocation rules see Moulin 1995).

Consider the following cost allocation rule:

$$\phi_i = \frac{h_i\lambda_i}{\eta_{\mathcal{N}}} + c\lambda_i + \gamma_i, \tag{13}$$

where

$$\gamma_i = \frac{h_i\lambda_i}{\sum_{i\in\mathcal{N}} h_i\lambda_i} c\eta_{\mathcal{N}} \quad \text{if} \quad \frac{\ln(\frac{1}{1-\alpha_{\mathcal{N}}})}{w_0} \leq \sqrt{\frac{\sum_{i\in\mathcal{N}} h_i\lambda_i}{c}} \tag{14}$$

and, otherwise (if $\frac{\ln(\frac{1}{1-\alpha_{\mathcal{N}}})}{w_0} > \sqrt{\frac{\sum_{i\in\mathcal{N}} h_i\lambda_i}{c}}$),

$$\gamma_i = \begin{cases} c\eta_{\mathcal{N}} - c\sqrt{\frac{\sum_{i\in\mathcal{N}, i\neq i_{max}} h_i\lambda_i}{c}} & \text{if} \quad i = i_{max}, \text{ and} \\ c\frac{h_i\lambda_i}{\sum_{i\in\mathcal{N}, i\neq i_{max}} h_i\lambda_i}\sqrt{\frac{\sum_{i\in\mathcal{N}, i\neq i_{max}} h_i\lambda_i}{c}} & \text{if} \quad i \neq i_{max}, \end{cases} \tag{15}$$

with again $i_{max} \in \{i : \alpha_i = \max(\alpha_1, ..., \alpha_n)\}$.

Remark 3. If the set $\{i : \alpha_i = \max(\alpha_1, ..., \alpha_n)\}$ has multiple indices, then we can arbitrarily choose any index in this set to be i_{max}. We can also let the firms with the highest service

level share the portion of the capacity cost $c\eta_{\mathcal{N}} - c\sqrt{\frac{\sum_{i\in\mathcal{N},i\neq i_{max}} h_i\lambda_i}{c}}$ in addition to the portion $c\frac{h_i\lambda_i}{\sum_{i\in\mathcal{N},i\neq i_{max}} h_i\lambda_i}\sqrt{\frac{\sum_{i\in\mathcal{N},i\neq i_{max}} h_i\lambda_i}{c}}$.

Under the above allocation rule, each firm (1) incurs its own delay cost, $\frac{h_i\lambda_i}{\eta_{\mathcal{N}}}$ and (2) a portion of total capacity cost, $c\lambda_i + \gamma_i$. The portion of total capacity cost has itself two parts: (a) an amount proportional to the firm's demand rate that can be directly attributed to each firm (this amount corresponds to the minimum cost needed to satisfy demand from this firm) and (b) a portion of the cost of buffer capacity. This portion is non-decreasing in the demand rate, delay cost, and service level of each firm. If $\frac{\ln(\frac{1}{1-\alpha_{\mathcal{N}}})}{w_0} \leq \sqrt{\frac{\sum_{i\in\mathcal{N}} h_i\lambda_i}{c}}$, this fraction is proportional to the firms' demand-weighted delay costs. If $\frac{\ln(\frac{1}{1-\alpha_{\mathcal{N}}})}{w_0} > \sqrt{\frac{\sum_{i\in\mathcal{N}} h_i\lambda_i}{c}}$ (the case where the service level constraint is more restrictive), firm i_{max} determines the service level requirement for the entire system. Therefore, it is treated differently to ensure that it is allocated a portion of the cost that is sufficiently high so that other firms do not break away from the coalition. This allocation appears to be consistent with those observed in practice, where combinations of volume based and capacity/service level based fees are common; see for example Gans and Zhou (2003, 2007) and Aksin et al. (2008).

Theorem 4.2 *The cost allocation rule $\phi = (\phi_1, ..., \phi_n)$ as specified in (13)-(15) is in the core. That is, under this cost allocation, no subset of the firms in $\mathcal{N}$ has an incentive to secede from the grand coalition.*

Remark 4. In general, a simple proportional cost allocation policy may not be in the core. For example, consider the case with pure service level constraints and with one of the firms requiring a much higher service level than the rest (the extreme case being only one of the firms requiring a service level). Then, clearly, all but one of the firms prefer not to join the grand coalition. Similarly, we can show that other common allocation schemes, such as the Shapley value, may not be in the core. In general, our cooperative game is not a concave game.

4.3 The Case of Waiting Time-based Service Levels

In this section, we consider the case where the service level is specified in terms of waiting time. Let $Q_{\mathcal{N}}$, a random variable, refer to customer waiting time in the shared facility. Then, the capacity

optimization problem can be stated as

$$\begin{aligned} &\text{Minimize } \ z_{\mathcal{N}}(\mu_{\mathcal{N}}) = c\mu_{\mathcal{N}} + \frac{\sum_{i\in\mathcal{N}} h_i\lambda_i}{\mu_{\mathcal{N}} - \sum_{i\in\mathcal{N}} \lambda_i} \\ &\text{subject to } P(Q_{\mathcal{N}} \le q_0) = 1 - \frac{\sum_{i\in\mathcal{N}} \lambda_i}{\mu_{\mathcal{N}}} e^{-(\mu_{\mathcal{N}} - \sum_{i\in\mathcal{N}} \lambda_i)q_0} \ge \alpha_{\mathcal{N}}, \\ &\qquad \sum_{i\in\mathcal{N}} \lambda_i/\mu_{\mathcal{N}} \le 1, \end{aligned} \tag{16}$$

where $\alpha_{\mathcal{N}} = \max(\alpha_1, ..., \alpha_n)$.

Similar to the case without capacity sharing, in the absence of service level constraints, the optimal capacity level is given by $\sum_{i\in\mathcal{N}} \lambda_i + \sqrt{\frac{\sum_{i\in\mathcal{N}} h_i\lambda_i}{c}}$. In systems with service level constraints but no delay costs, the optimal capacity level is given by the smallest $\mu_{\mathcal{N}}$ that satisfies the service level constraint or, equivalently, the optimal capacity is the solution to the following equation

$$\ln(\lambda_{\mathcal{N}} q_0) + \lambda_{\mathcal{N}} q_0 - \ln(1-\alpha_{\mathcal{N}}) = \ln(\mu_{\mathcal{N}} q_0) + \mu_{\mathcal{N}} q_0, \tag{17}$$

where $\lambda_{\mathcal{N}} = \sum_{i=1}^n \lambda_i$. Let $\bar{\mu}_{\mathcal{N}}(\alpha_{\mathcal{N}}, \lambda_{\mathcal{N}})$ be the solution to the above equation and $\bar{\eta}_{\mathcal{N}}(\alpha_{\mathcal{N}}, \lambda_{\mathcal{N}}) = \bar{\mu}_{\mathcal{N}}(\alpha_{\mathcal{N}}, \lambda_{\mathcal{N}}) - \lambda_{\mathcal{N}}$, the associated amount of buffer capacity (note that $\bar{\mu}_{\mathcal{N}} > \lambda_{\mathcal{N}}$). In contrast to the case where service level is specified in terms of delay, the amount of buffer capacity is not invariant to $\lambda_{\mathcal{N}}$ and is indeed increasing in $\lambda_{\mathcal{N}}$. This leads to the following result.

Theorem 4.3 *Capacity sharing may not be beneficial and it is possible for* $z^*_{\mathcal{N}} > z^*_{1,...,n}$ *and* $\mu^*_{\mathcal{N}} > \sum_{i=1}^n \mu^*_i$.

The above result can be proven using a counter-example. Consider the case where $h_i = 0$ for all i and $\alpha_1 > 0$ but $\alpha_i = 0$ for all $i \neq 1$. For the case without capacity sharing, we have $\bar{\eta}_i(0, \lambda_i) = 0$ for $i \neq 1$. However, $\eta_1(\alpha_1, \lambda_1) > 0$ and is given by the solution to

$$-\ln(1-\alpha_1) = \ln(1 + \frac{z}{\lambda_1}) + zq_0.$$

For the system with a single shared facility, we have $\bar{\eta}_{\mathcal{N}}(\alpha_1, \lambda_{\mathcal{N}})$, where $\bar{\eta}_{\mathcal{N}}(\alpha_1, \lambda_{\mathcal{N}})$ is the solution to

$$-\ln(1-\alpha_1) = \ln(1 + \frac{z}{\lambda_{\mathcal{N}}}) + zq_0.$$

Then, it is clear that $\bar{\eta}_{\mathcal{N}}(\alpha_1, \lambda_{\mathcal{N}}) > \bar{\eta}_1(\alpha_1, \lambda_1)$ and, consequently, $z_{\mathcal{N}} = c\bar{\eta}_{\mathcal{N}}(\alpha_1, \lambda_{\mathcal{N}}) + c\lambda_{\mathcal{N}} > \sum_{i\in\mathcal{N}} \bar{z}_i = c\bar{\eta}_1(\alpha_1, \lambda_1) + c\lambda_{\mathcal{N}}$. This is due to fact that buffer capacity is increasing in the demand rate, with the demand rates of different firms having marginal effects on the buffer capacity.

Hence, surprisingly, capacity sharing may not be even beneficial when service level constraints are imposed on waiting time in the queue even under the M/M/1 setting. This means that in contrast to the common sense that risk pooling is beneficial, capacity sharing should be taken with caution when the service level constraints are imposed on waiting time in the queue.

Notice that without service level constraints, the cooperative game can be shown as a submoduar game (which is due to that in this case the optimal cost $z^*_{\mathcal{N}} = c\sum_{i\in\mathcal{N}}\lambda_i + 2\sqrt{c\sum_{i\in\mathcal{N}}h_i\lambda_i},$) and the core exists. As we can see that the existence of the core for systems when the service level constraints are imposed on delay is due to the the buffer capacity for the service level constraints is independent of the total demand rate, i.e., when the service level constraints are active, the optimal buffer capacity level is determined by the firm with the highest service level only. In this case, capacity sharing always lowers the buffer capacity. However, when the service level constraints are imposed on waiting time, in general the core may not exist. This is due to that capacity sharing may lead to higher buffer capacity as we have shown in the above example since the optimal buffer capacity level may depend on both the service level constraints and the total demand rate. In particular, capacity sharing for a firm with high service level requirement but low demand rate and a firm with low service level requirement but high demand rate could be detrimental.

Although capacity sharing is not always beneficial in general, it is in the case where the firms have identical service level constraint and $\alpha_i = \alpha$ for all i.

Theorem 4.4 *If $\alpha_i = \alpha$, then $z^*_{\mathcal{N}} \le \bar{z}^*_{1,\ldots,n}$ and $\bar{\mu}^*_{\mathcal{N}} \le \sum_{i=1}^n \bar{\mu}^*_i$.*

Consider the following cost allocation scheme.

$$\phi_i = \frac{h_i\lambda_i}{\bar{\eta}_{\mathcal{N}}} + c\lambda_i + \gamma_i, \tag{18}$$

where

$$\gamma_i = \frac{h_i\lambda_i}{\sum_{i\in\mathcal{N}}h_i\lambda_i}c\sqrt{\frac{\sum_{i\in\mathcal{N}}h_i\lambda_i}{c}} \quad \text{if} \quad \bar{\eta}_{\mathcal{N}}(\alpha,\lambda_{\mathcal{N}}) \le \sqrt{\frac{\sum_{i\in\mathcal{N}}h_i\lambda_i}{c}} \tag{19}$$

and, otherwise (if $\bar{\eta}_{\mathcal{N}}(\alpha,\lambda_{\mathcal{N}}) > \sqrt{\frac{\sum_{i\in\mathcal{N}}h_i\lambda_i}{c}}$),

$$\gamma_i = \frac{\lambda_i}{\lambda_{\mathcal{N}}}c\bar{\eta}_{\mathcal{N}}(\alpha,\lambda_{\mathcal{N}}). \tag{20}$$

As stated in the following theorem, above allocation scheme is in the core.

Theorem 4.5 *If $\alpha_i = \alpha$ for all i, then the cost allocation rule $\phi = (\phi_1, ..., \phi_n)$ as specified in (18)-(20) is in the core. That is, under this cost allocation, no subset of the firms in $\mathcal{N}$ has an*

incentive to secede from the grand coalition.

In summary, if the service level constraints are imposed on total delay, then capacity sharing is always beneficial and we can identify a cost allocation scheme that is in the core. However, if the service level constraints are imposed on waiting times, then capacity sharing may not be beneficial because buffer capacity is no longer invariant to total demand. To our knowledge, this is a new result in the literature on capacity pooling in queueing systems. The result also suggests that capacity sharing should be used with caution and there might be settings where capacity sharing involving only a subset of the firms may be preferable to the grand coalition.

5 Systems with Service Priorities

We have so far assumed that customers in a shared facility, regardless of their firm affiliation, are served on a FCFS basis. This policy is simple to implement and evaluate and has the appearance of fairness. However, it is not system-optimal when there are multiple customer classes with different delay costs or different service level requirements. For example, for a system without service level requirements but with different delay costs for different customers classes, the so-called $c-\mu$ rule is known to be optimal; see, for instance, Jaiswal (1968) and Klimov (1974). Under the $c-\mu$ rule, customers are assigned priorities based on the product of their delay costs and their service rates (in our setting, this means that a higher service priority would be given to customers with higher delay costs). In the presence of service level requirements, the optimal policy is more complicated and must account for the interaction between delay costs and service levels, as well as other parameters.

In this section, we extend our analysis to settings where an optimal priority policy is used and investigate whether or not, under an optimal policy, there is a cost allocation that is in the core. The analysis of systems with priorities is notoriously difficult and, to our knowledge, there are no results in the literature regarding the nature of the optimal priority policy (in the presence of both delay costs and service level requirements) and no known closed form expressions for performance evaluation for systems that operate under such a policy. Also, to our knowledge, there are no results regarding the existence of the core for queueing systems that operate under a priority policy, optimal or otherwise. Other than assuming that a priority policy is used, the assumptions of the model we consider are the same as those of our original model described in Sections 3 and 4.

In order to analyze the associated cooperative game we resort to an indirect approach, the so-called achievable region method. Without loss of generality, we assume that $h_1 \geq h_2 \geq \cdots \geq h_n > 0$. We assume that the shared system operates under the optimal preemptive priority policy (the case

without preemption can be similarly analyzed). We assume that each firm is subject to a service level constraint on its expected delay (Unfortunately, other forms of service levels are substantially more difficult to analyze). The achievable region method considers the class of mixed preemptive priority policies. Note that for two strict preemptive priority policies $\mathcal{P}_1$ and $\mathcal{P}_2$, if at the beginning of each busy period, we use policy $\mathcal{P}_1$ with probability $1-\beta$ and policy $\mathcal{P}_2$ with probability β, the resulting policy is a mixed preemptive priority policy. Based on Coffman and Mitrani (1980), for coalition $\mathcal{J}$ given capacity μ (for $\mu > \lambda_{\mathcal{J}}, \lambda_{\mathcal{J}} = \sum_{i\in\mathcal{J}} \lambda_i$), the feasible region for the vector of the expected delay in $(E[W_i], i \in \mathcal{J})$ under any mixed preemptive priority policy can be described by the following polyhedron:

$$\begin{aligned} &\sum_{i\in\mathcal{V}} \lambda_i E[W_i] \geq \frac{\lambda_{\mathcal{V}}}{\mu - \lambda_{\mathcal{V}}}, \text{for all } \mathcal{V} \subseteq \mathcal{J}, \\ &E[W_i] \geq 0, \text{for all } i \in \mathcal{J}. \end{aligned} \tag{21}$$

Suppose that the service level constraint for each firm i is given by $E[W_i] \leq w_i$, i.e., firm i requires its expected delay should not be more than w_i. We can readily show that the problem of jointly deciding on the optimal capacity and the optimal priority order for coalition $\mathcal{J}$ is specified by the solution to the following optimization problem.

$$\begin{aligned} z^*_{\mathcal{J}} = &\min_{E[W_i], i\in\mathcal{J}; \mu} \sum_{i\in\mathcal{J}} h_i \lambda_i E[W_i] + c\mu \\ \text{subject to } &\sum_{i\in\mathcal{V}} \lambda_i E[W_i] \geq \frac{\lambda_{\mathcal{V}}}{\mu - \lambda_{\mathcal{V}}}, \text{for all } \mathcal{V} \subseteq \mathcal{J}, \\ &0 \leq E[W_i] \leq w_i, i \in \mathcal{J}, \\ &\mu > \lambda_{\mathcal{J}}. \end{aligned} \tag{22}$$

Let $E[W^P_{i,\mathcal{J}}], i \in \mathcal{J}$ and $\mu^P_{\mathcal{J}}$ be the optimal solution to (22). Noting that the above problem is a convex optimization problem, it can be reformulated into an equivalent problem using the Lagrangian method (see Bertsekas 1999). Let θ_i be the Lagrange multiplier for the service level constraint $E[W_i] \leq w_i$, for all $i \in \mathcal{J}$. Then, the optimal solution to the following problem is also

optimal for the original problem in (22):

$$\hat{z}^*_{\mathcal{J}} = \max_{\theta_i \geq 0, i \in \mathcal{J}} \min_{E[W_i], i \in \mathcal{J}, \mu} \sum_{i \in \mathcal{J}} (h_i + \theta_i)\lambda_i E[W_i] + c\mu - \sum_{i \in \mathcal{J}} \theta_i w_i \tag{23}$$

$$\text{subject to } \sum_{i \in \mathcal{V}} \lambda_i E[W_i] \geq \frac{\lambda_{\mathcal{V}}}{\mu - \lambda_{\mathcal{V}}}, \text{for all } \mathcal{V} \subseteq \mathcal{J},$$

$$E[W_i] \geq 0, \text{for all } i \in \mathcal{J},$$

$$\mu > \lambda_{\mathcal{J}}.$$

Based on Proposition 5.2.1 of Bertsekas (1999), we can show that $z^*_{\mathcal{J}} = \hat{z}^*_{\mathcal{J}}$, i.e., there is no duality gap. Note that the Lagrange multiplier θ_i can be viewed as the delay cost induced by the corresponding service level constraint for firm i. Hence, the presence of service level constraints might affect the optimal priority order. Let $\theta^*_{i,\mathcal{J}}$ be the optimal Lagrangian multiplier for firm i in coalition $\mathcal{J}$. The Lagrangian problem implies the following result.

Lemma 5.1 *The optimal priority policy for coalition $\mathcal{J}$ is a strict preemptive priority policy and the priority order is decreasing in $h_i + \theta^*_{i,\mathcal{J}}$, with the firm with the highest value of $h_i + \theta^*_{i,\mathcal{J}}$ having the highest priority.*

As we can see, in the presence of service level constraints, the $c-\mu$ rule may not be optimal and the optimal priority order is determined by the values $h_i + \theta^*_{i,\mathcal{J}}, i \in \mathcal{J}$. To the best of our knowledge, this is the first result in the literature to characterize the optimal priority policy in a system with both delays costs and service level constraints. Note that, although we do not have an explicit expression for the parameters $\theta^*_{i,\mathcal{J}}$, the $\theta^*_{i,\mathcal{J}}$s can be easily computed since the dual problem is a concave optimization problem (see Proposition 5.1.2 of Bertsekas 1999). Next, we show that the optimal cost in a shared system is submodular in the set of firms involved.

Theorem 5.2 *$z^*_{\mathcal{J}\cup\mathcal{T}} + z^*_{\mathcal{J}\cap\mathcal{T}} \leq z^*_{\mathcal{J}} + z^*_{\mathcal{T}}$ for all $\mathcal{J}, \mathcal{T} \subseteq \mathcal{N}$, i.e., the cooperative game is a submodular game.*

The above result is important because it is well known that a cooperative game that is submodular admits an allocation of total cost among that is in the core. In particular, the Shapley value is one such allocation (see Shapley, 1971 for a more detailed discussion on submodular games).

Under the Shapley value, firm i is allocated a fraction of total cost specified by

$$\phi_i = \sum_{\mathcal{J} \subseteq \mathcal{N} \setminus \{i\}} \frac{|\mathcal{J}|!(n - |\mathcal{J}| - 1)!}{n!} [z^*_{\mathcal{J}\cup\{i\}} - z^*_{\mathcal{J}}], i = 1, \cdots, n.$$

In order to show that the Shapley value allocates each firm its own delay cost and a fraction of capacity cost, we need to show ϕ_i can be expressed as

$$\phi_i = h_i \lambda_i E[W^P_{i,\mathcal{N}}] + \beta_i, \tag{24}$$

where the first term is the expected delay cost and β_i is the capacity cost for firm i, with $\beta_i \geq 0$.

Theorem 5.3 *The capacity sharing cost allocation scheme defined in (24) is in the core, with each firm incurring its own delay cost and a positive proportion of the total capacity cost, i.e.,* $\beta_i \geq 0, i \in \mathcal{N}$.

The proof follows from first noting that we can always express the Shapley value allocation ϕ_i as the sum of two terms: $h_i \lambda_i E[W^P_{i,\mathcal{N}}]$ (which is the optimal solution to (22) for the grand coalition $\mathcal{N}$) and β_i, where β_i can be either positive or negative. To do so, note that $\sum_{i=1}^n \beta_i = c\mu^P_{\mathcal{N}}$. This is because $z^*_{\mathcal{N}} = \sum_{i=1}^n \phi_i = \sum_{i=1}^n [h_i \lambda_i E[W^P_{i,\mathcal{N}}] + \beta_i] = \sum_{i=1}^n h_i \lambda_i E[W^P_{i,\mathcal{N}}] + c\mu^P_{\mathcal{N}}$. Also, note that, by virtue of the submodularity of the cooperative game we have $\sum_{i \in \mathcal{J}} \phi_i \leq z^*_{\mathcal{J}}$ for any subcoalition $\mathcal{J}$. In other words, any subset of the firms are better off receiving the cost allocation under the grand coalition than receiving the cost allocation in a subcoalition on their own. This also means that a firm that joins a subset of firms must make a positive payment toward capacity cost, in addition to incurring its own delay cost. Otherwise, the other firms would be better off without it (a firm that joins a sub-coalition increases the workload of the subcoalition; therefore, it either increases delay for everyone or increases capacity cost). Consequently, each member of any coalition must be responsible for a non-negative fraction of capacity cost or, equivalently, we must have $\beta_i > 0$, with $\beta_i = \phi_i - h_i \lambda_i E[W^P_{i,\mathcal{N}}]$.

We conclude this section by noting that our analysis can be extended to systems that operate under a non-preemptive priority policy. See for example Shanthikumar and Yao (1992) for a treatment of non-preemptive priority policies using the achievable region method; see also the end of the proof of Theorem 5.2 in the Appendix for more details.

6 Systems with Heterogeneous Work Contents and General Arrival and Service Times

In this section, we examine the impact of relaxing some of the assumptions we have made so far and consider (1) systems with heterogeneous work contents and (2) systems where the customer

inter-arrival times and processing times are not necessarily exponentially distributed. Our objective here is not to provide a comprehensive analysis, which is outside the scope of this paper, but rather to offer some insights and evaluate the extent to which results we obtained under these assumptions would continue to hold. More importantly, our objective is to highlight settings where capacity sharing may not be beneficial and/or settings where identifying a cost allocation that is in the core is not possible for systems under the FCFS policy. For such settings, we provide conditions for the stability of smaller sub-coalitions. Exact analysis for the general systems is difficult. Therefore, to obtain these insights, we rely throughout on approximations that have been extensively used in the literature. Throughout this section, we restrict our analyses to the case of pure delay costs, although some of the results can be extended to systems with service level constraints, e.g., for systems with service priorities.

6.1 Capacity Sharing among Firms with Heterogeneous Work Contents

In the models described in Sections 3, 4 and 5, we assumed that work contents are homogeneous across customers from different firms, so that processing times are identically distributed regardless of firm affiliation. In this section, we consider a system where the work contents associated with customers vary from firm to firm.

Customer processing time is still a random variable of the form Y_i/μ, for firm i, but Y_i is now exponentially distributed with mean l_i (recall that, in the original model, $l_i = 1$ for all $i \in \mathcal{N}$). Customers are still served on a FCFS basis. In the shared system, with n firms sharing a single facility, service times are no longer exponentially and identically distributed. Instead the distribution of service time X is a mixture of exponential distributions (i.e., the distribution is hyper-exponential) with mean $E(X) = \sum_{i\in\mathcal{N}} p_i l_i/\mu_{\mathcal{N}}$ and second moment $E(X^2) = 2\sum_{i\in\mathcal{N}} p_i l_i^2/\mu_{\mathcal{N}}^2$, where $p_i = \lambda_i/\sum_{i\in\mathcal{N}}\lambda_i$. The resulting squared coefficient of variation $C_s^2 = \frac{E(X^2)-E(X)^2}{E(X)^2}$ is strictly greater than 1 ($C_s^2 = 1$ is the case where firms have identical work contents) and increases with the differences in the amount of work content of different firms.

Without capacity sharing, the problem faced by firm i can be stated as

$$\begin{aligned} \text{minimize} \quad & z_i(\mu_i) = c\mu_i + \frac{h_i\lambda_i}{\mu_i/l_i - \lambda_i} \\ \text{subject to} \quad & \lambda_i l_i/\mu_i \leq 1, \end{aligned}$$

whereas with capacity sharing under the FCFS policy, the problem can be stated as

$$\begin{aligned}&\text{minimize} \quad z_{\mathcal{N}}(\mu_{\mathcal{N}}) = \sum_{i\in\mathcal{N}} h_i\lambda_i \left[\frac{(1+C_s^2)E(X)}{2(1-\rho)} + \frac{l_i}{\mu_{\mathcal{N}}}\right] + c\mu_{\mathcal{N}}\\ &\text{subject to } \rho = \frac{\sum_{i\in\mathcal{N}}\lambda_i l_i}{\mu_{\mathcal{N}}} \leq 1.\end{aligned}$$

The optimization problems in both cases are convex and therefore can be easily solved numerically. However, a closed form analytical solution is difficult to obtain in the case of the shared system because the shared facility no longer behaves like an M/M/1 queue and, instead, behaves like an M/G/1 queue. Note that expected delay now depends on both the mean and the variance of service time via the coefficient of variation C_s.

In Appendix B, we provide a numerical example showing that capacity sharing may not be optimal in this case. Moreover, we also show that sub-coalitons could be optimal in this case. In the following proposition, we provide a sufficient condition for capacity sharing to be beneficial.

Theorem 6.1 *Capacity sharing is beneficial if*

$$(\frac{h_i}{l_i} - \frac{h_j}{l_j})(l_i - l_j) \geq 0, \forall i \neq j. \tag{25}$$

In particular, if $h_i/l_i = h_j/l_j$ for all i, j, then the the capacity cost allocation scheme $c\lambda_i + \frac{h_i\lambda_i l_i}{\sum_{k=1}^n h_k\lambda_k l_k}\sqrt{\frac{\sum_{k=1}^n h_k\lambda_k l_k}{c}}$ for firm i is in the core.

To our knowledge, the above result is new to the literature. The result states that for capacity sharing to be beneficial a firm with a larger (smaller) work content should also have higher (lower) value of $\frac{h_i}{l_i}$, i.e., the delay cost h_i should increase at least proportionally to l_i. This is consistent the empirical observations that the more valuable of the service is, the longer the customer is willing to wait in Maister (1985). We note that Whitt (1999) also shows that pooling demand streams with different characteristics could be harmful. He shows this for settings without capacity optimization. van Dijk and Sluis (2008) also show that pooling two M/M/1 systems into an M/M/2 system may not be beneficial if the mean service times for two M/M/1 systems are different. They also provide a sufficient condition under which pooling is beneficial.

The case such that $h_i/l_i = h_j/l_j$ for all i, j corresponds to settings where delay costs are proportional to work content, so that firms that bring more work to the system per customer are also those that are more sensitive to delay. This prevents situations where a firm with a small work content but high delay cost leads the shared facility to invest in a large amount of capacity to

mitigate the delay cost of that firm. In that case, the other firms may do better by excluding the high delay cost/low work content firm from the coalition. Note that the condition in the theorem is independent of the demand rates. This points to the dominant role that differences in work content and delay costs play in determining whether or not capacity sharing is beneficial. When the core does not exist, then the above result suggests we shall pool firms with similar $\frac{h_i}{l_i}$ together, which is also observed in Figure 5.

6.2 Capacity Sharing among Firms with General Demand and Processing Times

We consider systems where customer inter-arrival times for each firm $i \in \mathcal{N}$ are independent and identically distributed (i.e., arrivals form a renewal process) with mean $1/\lambda_i$ and coefficient of variation C_{a_i}. Customer processing times are independent, identically distributed, and described by a random variable of the form Y/μ, where Y has a mean equal to one and coefficient of variation C_s. The parameter μ is again a scaling factor that corresponds to the service rate. Exact analysis is difficult for the system under a FCFS policy. Therefore, to obtain insights we rely on approximations that have been extensively used in the literature.

In systems without capacity sharing, each independent facility can thus be modeled as a $GI/G/1$ queue. To obtain an explicit expression for the expected delay cost, we rely on an approximation that is asymptotically correct when the demand rates are high (i.e., when $\lambda_i \to \infty$). In particular, we approximate the expected number of customers at firm i, given capacity level μ_i, as follows

$$E[L_i(\mu_i)] \approx \sigma_i^2 \frac{\lambda_i}{\mu_i - \lambda_i}, \tag{26}$$

where $\sigma_i = \sqrt{\frac{C_{a_i}^2 + C_s^2}{2}}$. Motivation and supporting arguments for the approximation can be found in Harrison (1985) and more recently in Bassombo et al. (2012) and the references therein. The problem faced by each firm can then be restated as

$$\text{Minimize } z_i(\mu_i) \approx \frac{h_i \lambda_i}{\mu_i - \lambda_i} \sigma_i^2 + c\mu_i,$$

subject to $\lambda_i/\mu_i \leq 1$. This leads to an optimal capacity given by $\mu_i^* = \lambda_i + \sigma_i \sqrt{\frac{h_i \lambda_i}{c}}$, and corresponding optimal cost

$$z_i^* = c\lambda_i + 2\sigma_i \sqrt{c h_i \lambda_i}. \tag{27}$$

Bassombo et al. (2012) show that this capacity is asymptotically optimal when the demand rate is high (i.e., $\lambda_i \to \infty$). It also reduces to the optimal capacity for the $M/M/1$ case (in that case, $\sigma_i = 1$). Note that the above expressions capture now explicitly the effect of both demand and processing time variability.

For systems with capacity sharing, the analysis is more complicated since the superposition of renewal processes is not necessarily a renewal process. To handle this difficulty, we approximate superposed renewal processes by a renewal process whose coefficient of variation is obtained via a two-moment approximation; see Albin (1984) and Whitt (1982). In particular, we approximate the arrival process to a facility shared by the N firms by a renewal process with rate $\sum_{i\in\mathcal{N}} \lambda_i$ and coefficient of variation $C_{a_\mathcal{N}}^2 = \sum_{i\in\mathcal{N}} \frac{\lambda_i {C_{a_i}}^2}{\sum_{i=1}^n \lambda_i}$. The capacity optimization problem can be stated as

$$\text{Minimize} \;\; z_\mathcal{N}(\mu_\mathcal{N}) \approx c\mu_\mathcal{N} + \sigma_\mathcal{N}^2 \frac{\sum_{i\in\mathcal{N}} h_i\lambda_i}{\mu_\mathcal{N} - \sum_{i\in\mathcal{N}} \lambda_i},$$

subject to $\sum_{i\in\mathcal{N}} \lambda_i/\mu_\mathcal{N} \leq 1$, where $\sigma_\mathcal{N} = \sqrt{\frac{C_{a_\mathcal{N}}^2 + C_s^2}{2}}$. Hence, the optimal capacity is given by $\mu_\mathcal{N}^* = \sum_{i\in\mathcal{N}} \lambda_i + \sigma_\mathcal{N}\sqrt{\frac{\sum_{i\in\mathcal{N}} h_i\lambda_i}{c}}$ and the optimal cost is given by

$$z_\mathcal{N}^* = c\sum_{i\in\mathcal{N}} \lambda_i + 2\sigma_\mathcal{N}\sqrt{c\sum_{i\in\mathcal{N}} h_i\lambda_i}. \tag{28}$$

Observation 6.2 *Capacity sharing can lead to higher total cost in the system. That is, it is possible to have* $z_\mathcal{N}^* > \sum_{i\in\mathcal{N}} z_i^*$.

To see why the observation is true, consider an example with two firms, where firm 1 has negligible unit delay cost ($h_1 \approx 0$) but high variability ($\sigma_1 \gg 0$) while firm 2 has high unit delay cost ($h_2 \gg 0$) but negligible variability ($\sigma \approx 0$). Then, without capacity sharing, each firm would invest in negligible buffer capacity as either the unit delay cost is negligible or variability is negligible. When the firms share the same facility, the overall variability is positive. Therefore, there is congestion which, in turn, leads to delay costs being incurred by customers of firm 2, requiring an investment in buffer capacity.

We note that Whitt (1999) also shows that pooling demand streams with different characteristics could be harmful. He shows this for settings without capacity optimization. Our results show that this effect is also present when capacity is optimized. We also show it for the specific setting of Section 6.1 where variability is driven by differences in work contents.

Although capacity sharing is not always beneficial, it is still possible to identify plausible ranges

of parameter values for which capacity sharing is beneficial. A necessary and sufficient condition for capacity sharing to be preferable is given by

$$\sigma_{\mathcal{N}} \sqrt{\sum_{i=1}^{n} h_i \lambda_i} \leq \sum_{i \in \mathcal{N}} \sigma_i \sqrt{h_i \lambda_i}, \tag{29}$$

which follows from a direct comparison of the total buffer capacity cost in systems with and without capacity sharing. A sufficient condition for the inequality in (29) to hold is for

$$\frac{\sigma_{\mathcal{N}}}{\sigma_i} \leq \frac{\sum_{i \in \mathcal{N}} \sqrt{h_i \lambda_i}}{\sqrt{\sum_{i=1}^{n} h_i \lambda_i}} \quad \text{for all} \quad i \in \mathcal{N}.$$

In other words, capacity sharing is beneficial if the variability associated with capacity sharing among all the firms does not exceed a certain threshold relative to the variability associated with firms investing in their own independent facilities. The following theorem provides a more explicit sufficient condition under which capacity sharing is beneficial.

Theorem 6.3 *Capacity sharing is beneficial if, for each pair of firms i and j, $(h_i - h_j)(\sigma_i - \sigma_j) \geq 0$. In other words, capacity sharing is beneficial if firms with higher delay costs also have higher demand variability. In particular, the cost allocation defined by*

$$\phi_i = \frac{h_i \lambda_i}{\sqrt{\frac{\sum_{k=1}^{n} h_k \lambda_k}{c}}} + c\lambda_i + \frac{h_i \lambda_i}{\sum_{k=1}^{n} h_k \lambda_k} \sigma_{\mathcal{N}} \sqrt{\frac{\sum_{k=1}^{n} h_k \lambda_k}{c}} \tag{30}$$

is in the core if $\sigma_{\mathcal{N}}^2 / \sigma_{\mathcal{J}}^2 \leq \gamma_{\mathcal{J}}$ where $\gamma_{\mathcal{J}} = \frac{\sum_{k \in \mathcal{N}} h_k \lambda_k}{\sum_{k \in \mathcal{J}} h_k \lambda_k}$ for all $\mathcal{J} \subseteq \mathcal{N}$.

To our knowledge, the condition in Theorem 6.3 is also new to the literature. Note that the variability parameter here is affected by differences in inter-arrival time variability. Hence, this result complements the result of section 6.1 where the variability parameter was determined by differences in work contents. The condition is independent of the firms' demand rates. In particular, if $\sigma_i = \sigma_j$ for all i and j, then capacity sharing is always beneficial. The case of exponential inter-arrival times and processing times of course satisfies this condition, but it is also satisfied by a broader class of problems.

We conclude this section by noting that for systems that operate under the FCFS policy, in addition to highlighting conditions under which the core is empty, the results presented here provide conditions under which capacity sharing continues to be beneficial. These conditions can be used to identify settings where stable sub-coalitions might arise. They can also be used to guide decisions

for whether or not individual firms should invest in shared facilities with specific firms.

7 Summary and Concluding Comments

In this paper, we studied the benefit of capacity sharing among independent firms when capacity is optimized to take into account the costs and service level requirements of the different firms. We characterized settings under which capacity sharing is beneficial and proposed cost allocation schemes that guarantee that the grand coalition is stable for systems operate under either a FCFS policy or an optimal priority policy. However, for some systems that operate under a FCFS policy, we found that capacity sharing may not be beneficial. This is the case when service levels are specified in terms of waiting time instead of total delay and when the firms are heterogeneous in their characteristics, including work contents, service time variability, and delay costs. For these cases, we characterize conditions, under which capacity sharing may still be beneficial for a subset of the firms. These conditions provide insights into the characteristics of firms that would benefit from sharing capacity. In Appendix B, we also numerically explored the benefits of capacity sharing under different system parameters and obtained some insights correspondingly.

There are various avenues for future research. It would be useful to generalize the analysis to systems where capacity sharing involves the sharing of a network of facilities whose capacities can be individually optimized. It would also be useful to treat more general cost functions, including those that may exhibit economies or diseconomies of scale. Finally, it would be useful to consider systems where capacities can be varied only in discrete amounts. We expect the treatment of each of these cases to be be considerable more difficult and to require perhaps the development of alternative approaches or approximations.

Acknowledgments: The authors are grateful to the review team for their excellent suggestions that have improved this paper substantially.

References

O. Z. Aksin, F. Karaesmen, and E. L. Ormeci, "A Review of Workforce Cross-Training in Call Centers from an Operations Management Perspective," in *Workforce Cross Training Handbook*, D. Nembhard (editors), CRC Press, 211-240, 2007.

O.Z. Aksin, F. de Vericourt, and F. Karaesmen, "Call Center Outsourcing Contract Design and Choice," *Management Science*, **54**, 354-368 , 2008.

S. L. Albin, "Approximating a Point Process by a Renewal Process, II: Superposition Arrival Processes to Queues," *Operations Research*, **32**, 1133-1162, 1984.

G. Allon and A. Federgruen, "Competition in Service Industries," *Operations Research*, **55**, 37-55, 2007.

G. Allon and A. Federgruen, " Service Competition with General Queueing Facilities," *Operations Research*, **56**, 827-849, 2008.

S. Anily and M. Haviv, "Cooperation in Service Systems," *Operations Research*, **58**, 660-673, 2010.

S. Anily and M. Haviv, "Cost-allocation for the first order interaction joint replenishment model," *Operations Research*, **55**, 292-302, 2007.

A. Bassombo, Randhawa,R. S. and J. A. van Mieghem, " A Little Flexibility Is All You Need: On the Asymptotic Value of Flexible Capacity in Parallel Queuing Systems," *Operations Research*, **60**,1423-1435, 2012.

S. Benjaafar, "Performance Bounds for the Effectiveness of Pooling in Multi-Processing Systems," *European Journal of Operational Research*, **87**, 375-388, 1995.

S. Benjaafar, W. L. Cooper and J. S. Kim, "On the Benefits of Pooling in Production-Inventory Systems," *Management Science*, **51**, 548-565, 2005.

S. Benjaafar, E. Elahi and K. Donohue, "Outsourcing via Service Quality Competition," *Management Science*, **53**, 241-259, 2007.

N. Ben-Zvi and Y. Gerchak, "Inventory Centralization When Shortage Costs Differ: Priorities and Costs Allocation," working paper, Tel-Aviv University, 2005.

D. Bertsekas, *Nonlinear Programming*, Athena Scientific, 1999.

J. A. Buzacott, "Commonalities in Reengineered Business Processes: Models and Issues," *Management Science*, **42**, 768-782, 1996.

G. Cachon and P. Harker, "Competition and Outsourcing with Scale Economies," *Management Science*, **48**, 1314-1333, 2002.

G. Cachon and F. Zhang, "Obtaining Fast Service in a Queuing system via Performance-Based Allocation of Demand," *Management Science*, **53**, 408-420, 2007.

X. Chen and J. Zhang, "Duality Approaches to Economic Lot Sizing Games," working paper, New York University, 2006.

X. Chen and J. Zhang, "A Stochastic Programming Duality Approach to Inventory Centralization Games," *Operations Research*, **57**, 840-851, 2009.

S. Dewan and H. Mendelson,"User Delay Costs and Internal Pricing for a Service Facility," *Management Science*, **36**, 1502-1517, 1990.

N. M. van Dijk and E. Sluis, "To pool or not to pool in call centers," *Production and Operations Management*, 17(3) 296?305, 2008.

M. Dror and B. Hartman, "Shipment Consolidation: Who Pays for It and How Much?" *Management Science*, **53**, 7887, 2007.

N. Gans and Y-P. Zhou, "A Call-Routing Problem with Service-Level Constraints," *Operations Research*, **51**, 255-271, 2003.

N. Gans and Y-P. Zhou. "Call-Routing Schemes for Call-Center Outsourcing," to appear in *Manufacturing and Service Operations Management*, **9**, 33-50, 2007.

M. D. Garcia-Sanz, F. R. Fernandez, M. G. Fiestras-Janeiro, I. Garcia-Jurado and J. Puerto, "Cooperation in Markovian Queueing Models," *European Journal of Operational Research*, **188**, 485-495, 2008.

P. Gonzalez and C. Herrero, "Optimal Sharing of Surgical Costs in the Presence of Queues," *Mathematical Methods of Operations Research*, **59**, 435-446, 2004.

S. Gurumurthi and S. Benjaafar, "Modeling and Analysis of Flexible Queueing Systems," *Naval Research Logistics*, **51**, 755-782, 2004.

A. Y. Ha, "Optimal Pricing That Coordinates Queues with Customer-Chosen Service Requirements," *Management Science*, **47**, 915-930, 2001.

E. Hanany and Y. Gerchak, "Nash Bargaining over Inventory and Pooling Contracts," forthcoming in *Naval Research Logistics*, **55**, 541-550, 2008.

J. M. Harrison, *Brownian Motion and Stochastic Flow Systems*, John Wiley, New York, New York, 1985.

R. Hassin and M. Haviv, *To Queue or not to Queue*, Kluwer, Boston, 2003.

W. J. Hopp, E. Tekin and M. P. Van Oyen, "Benefits of Skill Chaining in Production Lines with Cross-Trained Workers," *Management Science*, **50**, 83-98, 2004.

S. M. Iravani, M. P. Van Oyen and K.T. Sims (2005). "Structural Flexibility: A New Perspective on the Design of Manufacturing and Service Operations," *Management Science*, **51**, 151-166.

N. K. Jaiswal, *Priority Queues*, New York, Academic Press, 1968.

O. Jouini, Y. Dallery and R. Nait-Abdallah, "Analysis of the Impact of Team-Based Organizations in Call Center Management," *Management Science*, 400-414, 2008.

W. C. Jordan, R.R. Inman and D. E. Blumenfeld, "Chained Cross-Training of Workers for Robust Performance," *IIE Transactions*, **36**, 953-967, 2004.

E. Kalai, M. I. Kamien and M. Rubinovitch, "Optimal Service Speeds in a Competitive Environment," *Management Science*,38, 1154-1163, 1992.

A. Katta and J. Sethuraman, "Cooperation in Queues, Working Paper, Columbia University, 2006.

E. Kemahlioglu-Ziya, "Formal Methods of Value Sharing in Supply Chains", Ph.D Thesis, Georgia Institute of Technology, 2004.

L. Kleinrock, *Queueing Systems, Computer Applications, Volume 2*, John Wiley & Sons, 1975.

G. P. Klimov, "Time Sharing Service Systems I.," *Theory of Probability and Its Application*, **19**, 532-551, 1974.

G. Koole and A. Pot, "An Overview of Routing and Staffing in Multi-Skill Customer Contact Centers," working paper, Vrije Universiteit, 2005.

D. H. Maister, "The psychology of waiting lines," in The Service Encounter: Managing Employee/Customer Interaction in Service Business, by John A. Czepiel, Michael R. Solomon, Carol F. Suprenant (Editor). Lexington Books, 1985

A. Mandelbaum, and M. I. Reiman, "On Pooling in Queueing Networks," *Management Science*, **44**, 971-981, 1998.

F. Maniquet, "A Characterization of the Shapley Value in Queueing Problems", *Journal of Economic Theory*, **109**, 90-103, 2003.

P. Milgrom and I. Segal, "Envelope Theorems for Arbitrary Choice Sets", *Econometrica*, , **70**, 583-601, 2002.

H. Mendelson and S. Whang, "Optimal Incentive-Compatible Priority Pricing for the M/M/1 Queue," *Operations Research*, **38**, 870-883, 1990.

H. Moulin, *Cooperative Microeconomics: a Game-Theoretic Introduction*, Princeton University Press, 1995.

H. Moulin and R. Strong, "Fair Queuing and Other Probabilistic Allocation Methods," *Mathematics of Operations Research*, **27**, 1-30, 2002.

A. Muller, M. Scarsini and M. Shaked, "The Newsvendor Game Has a Nonempty Core," *Games and Economic Behavior*, **38**, 118-126, 2002.

M. Nagarajan and G. Sošić, "Game-Theoretic Analysis of Cooperation Among Supply Chain Agents: Review and Extensions," *European Journal of Operational Research*, **187**, 719-745, 2008.

M. I. Reiman, "Open Queueing Networks in Heavy Traffic," *Mathematics of Operations Research*, 441-458, 1984.

L.S. Shapley, "Cores of convex games", *International Journal of Game Theory*, 1(1), 1126, 1971.

M. Sheikhzadeh, S. Benjaafar and D. Gupta, "Machine Sharing in Manufacturing Systems: Flexibility versus Chaining," *International Journal of Flexible Manufacturing Systems*, **10**, 351-378, 1998.

D. R. Smith and W. Whitt, "Resource Sharing for Efficiency in Traffic Systems," *The Bell System Technical Journal*, **60**, 1981.

S. Stidham, "On the Optimality of Single-Server Queueing Systems," *Operations Research*, **18**, 708-732, 1970.

S. Stidham, "Pricing and Capacity Decisions for a Service Facility: Stability and Multiple Local Optima," *Management Science*, **38**, 1121-1139, 1992.

E. Tekin, W. Hopp and M. V. Oyen, "Pooling Strategies for Call Center Agent Crosstraining," *IIE Transactions*, 41, 546-561, 2009.

W. van den Heuvel, P.E.M. Borm and H. Hamers, "Economic Lot-Sizing Games," *European Journal of Operational Research*, **176**, 1117-1130, 2007.

R. B. Wallace and W. Whitt, "A Staffing Algorithm for Call Centers with Skill-Based Routing," *Manufacturing and Service Operations Management*, **7**, 276-294, 2005.

W. Whitt, "The Queueing Network Analyzer," *The Bell System Technical Journal*, **62**, 2779-2815, 1983.

W. Whitt, "Approximating a Point Process by a Renewal Process, I: Two Basic Methods," *Operations Research*, **30**, 125-147, 1982.

W. Whitt, "Partitioning customers into service groups," *Management Science*, 45(11), 1999.

P. H. Zipkin, *Foundation of Inventory Management*, McGraw-Hill, 2000.

Appendices

Appendix A

Proof of Lemma 3.1: Noting that $\bar{\eta}_i(\alpha_i, \lambda_i)$ is the solution to

$$-\ln(1-\alpha_i) = \ln((\lambda_i + z)q_0) + (\lambda_i + z)q_0 - \ln(\lambda_i q_0) - \lambda_i q_0,$$

or equivalently

$$-\ln(1-\alpha_i) = \ln(1 + \frac{z}{\lambda_i}) + zq_0,$$

it is to verify that $\bar{\eta}_i(\alpha_i, \lambda_i)$ is nondecreasing in λ_i for all $\lambda_i > 0$.

We show that $\bar{\eta}_i(\alpha_i, \lambda_i)/\lambda_i$ is nonincreasing in λ_i by contradiction. Suppose that $\bar{\eta}_i(\alpha_i, x)/x$ is strictly increasing in x in the neighborhood of λ_i for some $\lambda_i > 0$, i.e., $\bar{\eta}_i(\alpha_i, \lambda_i + \delta)/(\lambda_i + \delta) > \bar{\eta}_i(\alpha_i, \lambda_i)/\lambda_i$ for any small enough $\delta > 0$. Then we must have

$$\ln(1 + \frac{\bar{\eta}_i(\alpha_i, \lambda_i + \delta)}{\lambda_i + \delta}) + (\lambda_i + \delta)\frac{\bar{\eta}_i(\alpha_i, \lambda_i + \delta)}{\lambda_i + \delta}q_0 > \ln(1 + \frac{\bar{\eta}_i(\alpha_i, \lambda_i)}{\lambda_i}) + \lambda_i \frac{\bar{\eta}_i(\alpha_i, \lambda_i)}{\lambda_i}q_0 = -\ln(1-\alpha_i)$$

for all $\delta > 0$ since $\ln(1+x)$ is an increasing function, which contradicts with the fact that

$$\ln(1 + \frac{\bar{\eta}_i(\alpha_i, \lambda_i + \delta)}{\lambda_i + \delta}) + (\lambda_i + \delta)\frac{\bar{\eta}_i(\alpha_i, \lambda_i + \delta)}{\lambda_i + \delta}q_0 = -\ln(1-\alpha_i).$$

Hence, $\bar{\eta}_i(\alpha_i, \lambda_i)/\lambda_i$ is nonincreasing in λ_i. ■

Proof of Theorem 4.1: To prove that $\sum_{i\in\mathcal{N}} \mu_i^* \geq \mu_{\mathcal{N}}^*$, note that

$$\begin{aligned}\sum_{i\in\mathcal{N}} \mu_i^* &= \sum_{i\in\mathcal{N}} \max\{\frac{\ln(\frac{1}{1-\alpha_i})}{w_0}, \sqrt{\frac{h_i\lambda_i}{c}}\} \geq \max\{\frac{\ln(\frac{1}{1-\alpha_{i_{max}}})}{w_0}, \sqrt{\frac{h_{i_{max}}\lambda_{i_{max}}}{c}}\} + \sum_{i\in\mathcal{N}, i\neq i_{max}} \sqrt{\frac{h_i\lambda_i}{c}} \\ &\geq \max\{\frac{\ln(\frac{1}{1-\alpha_{\mathcal{N}}})}{w_0}, \sqrt{\frac{\sum_{i\in\mathcal{N}} h_i\lambda_i}{c}}\} = \mu_{\mathcal{N}}^*,\end{aligned}$$

where $i_{max} \in \{i : \alpha_i = \max(\alpha_1, ..., \alpha_n)\}$. The last inequality is due to that $\sqrt{\frac{h_{i_{max}}\lambda_{i_{max}}}{c}} + \sum_{i\in\mathcal{N}, i\neq i_{max}} \sqrt{\frac{h_i\lambda_i}{c}} \geq \sqrt{\frac{\sum_{i\in\mathcal{N}} h_i\lambda_i}{c}}$. In order to prove that $z_{\mathcal{N}}^* \leq z_{1,...,n}^*$, we distinguish two cases.

(1) $\frac{\ln(\frac{1}{1-\alpha_{\mathcal{N}}})}{w_0} \leq \sqrt{\frac{\sum_{i\in\mathcal{N}} h_i\lambda_i}{c}}$: In this case, we have

$$\begin{aligned}z_{\mathcal{N}}^* &= \frac{\sum_{i\in\mathcal{N}} h_i\lambda_i}{\sqrt{\frac{\sum_{i\in\mathcal{N}} h_i\lambda_i}{c}}} + c\sum_{i\in\mathcal{N}} \lambda_i + c\sqrt{\frac{\sum_{i\in\mathcal{N}} h_i\lambda_i}{c}} \leq \sum_{i\in\mathcal{N}} (\frac{h_i\lambda_i}{\sqrt{\frac{h_i\lambda_i}{c}}} + c\lambda_i + c\sqrt{\frac{h_i\lambda_i}{c}}) \\ &\leq \sum_{i\in\mathcal{N}} (c(\lambda_i + \eta_i) + \frac{h_i\lambda_i}{\eta_i}) \leq z_{1,...,n}^*.\end{aligned}$$

The first inequality is due to that $\sqrt{\sum_{i\in\mathcal{N}} h_i\lambda_i c} \leq \sum_{i\in\mathcal{N}} \sqrt{h_i\lambda_i c}$. The second inequality is due to that the presence of the service level constraints would not decrease the optimal costs for firms.

(2) $\frac{\ln(\frac{1}{1-\alpha_{\mathcal{N}}})}{w_0} > \sqrt{\frac{\sum_{i\in\mathcal{N}} h_i\lambda_i}{c}}$: In this case, we have

$$\begin{aligned} z^*_{\mathcal{N}} =& \frac{\sum_{i\in\mathcal{N}} h_i\lambda_i}{\frac{\ln(\frac{1}{1-\alpha_{\mathcal{N}}})}{w_0}} + c\sum_{i\in\mathcal{N}}\lambda_i + c\frac{\ln(\frac{1}{1-\alpha_{\mathcal{N}}})}{w_0} \\ \leq& \frac{h_{i_{max}}\lambda_{i_{max}}}{\frac{\ln(\frac{1}{1-\alpha_{i_{max}}})}{w_0}} + c\lambda_{i_{max}} + c\frac{\ln(\frac{1}{1-\alpha_{i_{max}}})}{w_0} + \sum_{i\in\mathcal{N}, i\neq i_{max}} (\frac{h_i\lambda_i}{\sqrt{\frac{h_i\lambda_i}{c}}} + c\lambda_i + c\sqrt{\frac{h_i\lambda_i}{c}}) \\ \leq& \sum_{i\in\mathcal{N}} (c(\lambda_i+\eta_i) + \frac{h_i\lambda_i}{\eta_i}) \leq z^*_{1,\ldots,n}. \end{aligned}$$

The first inequality is due to that $\alpha_{i_{max}} = \alpha_{\mathcal{N}}$. The second inequality is due to that $\eta_{i_{max}} = \frac{\ln(\frac{1}{1-\alpha_{\mathcal{N}}})}{w_0}$ and the presence of the service level constraints would not decrease the optimal costs for firms.

■

Proof of Theorem 4.2: We distinguish two cases here.

(1) $\frac{\ln(\frac{1}{1-\alpha_{\mathcal{N}}})}{w_0} \leq \sqrt{\frac{\sum_{i\in\mathcal{N}} h_i\lambda_i}{c}}$. First note that $z^*_{\mathcal{J}} \geq c\sum_{i\in\mathcal{J}}\lambda_i + 2\sqrt{c\sum_{i\in\mathcal{J}} h_i\lambda_i}$. Since

$$\sum_{i\in\mathcal{J}}\phi_i - \left[c\sum_{i\in\mathcal{J}}\lambda_i + 2\sqrt{c\sum_{i\in\mathcal{J}} h_i\lambda_i}\right] = 2\left[\sum_{i\in\mathcal{J}} h_i\lambda_i\sqrt{\frac{c}{\sum_{i\in\mathcal{N}} h_i\lambda_i}} - \sqrt{c\sum_{i\in\mathcal{J}} h_i\lambda_i}\right] \leq 0,$$

we have $z^*_{\mathcal{J}} \geq \sum_{i\in\mathcal{J}}\phi_i$, $\forall\mathcal{J}\subseteq\mathcal{N}$. It follows that the allocation rule is in the core.

(2) $\frac{\ln(\frac{1}{1-\alpha_{\mathcal{N}}})}{w_0} > \sqrt{\frac{\sum_{i\in\mathcal{N}} h_i\lambda_i}{c}}$. For coalition $\mathcal{J}\subseteq\mathcal{N}\setminus\{i_{max}\}$, we have $z^*_{\mathcal{J}} \geq c\sum_{i\in\mathcal{J}}\lambda_i + 2\sqrt{c\sum_{i\in\mathcal{J}} h_i\lambda_i}$. Since $c\sum_{i\in\mathcal{J}}\lambda_i + 2\sqrt{c\sum_{i\in\mathcal{J}} h_i\lambda_i} \geq \sum_{i\in\mathcal{J}}\left[\frac{h_i\lambda_i}{\sqrt{\frac{\sum_{i\in\mathcal{N}, i\neq i_{max}} h_i\lambda_i}{c}}} + c\lambda_i + c\frac{h_i\lambda_i}{\sum_{i\in\mathcal{N}, i\neq i_{max}} h_i\lambda_i}\sqrt{\frac{\sum_{i\in\mathcal{N}, i\neq i_{max}} h_i\lambda_i}{c}}\right]$ $\geq \sum_{i\in\mathcal{J}}\phi_i$, we have $z^*_{\mathcal{J}} \geq \sum_{i\in\mathcal{J}}\phi_i$, $\forall\mathcal{J}\subseteq\mathcal{N}\setminus\{i_{max}\}$. If $i_{max}\in\mathcal{J}$, then $z^*_{\mathcal{J}} = \frac{\sum_{i\in\mathcal{J}} h_i\lambda_i}{\frac{\ln(\frac{1}{1-\alpha_{\mathcal{N}}})}{w_0}} + c\sum_{i\in\mathcal{J}}\lambda_i + c\frac{\ln(\frac{1}{1-\alpha_{\mathcal{N}}})}{w_0} \geq \sum_{i\in\mathcal{J}}\phi_i, \forall\mathcal{J} : i_{max}\in\mathcal{J}$ (Note that if $i_{max}\in\mathcal{J}$, then the optimal capacity for coalition $\mathcal{J}$ is given by $\sum_{i\in\mathcal{J}}\lambda_i + \frac{\ln(\frac{1}{1-\alpha_{\mathcal{N}}})}{w_0}$). Consequently, the allocation is in the core.

■

Proof of Theorem 4.4: Noting that $\bar{\eta}_i(\alpha,\lambda_i)$ is the solution to

$$-\ln(1-\alpha) = \ln(1+\frac{z}{\lambda_i}) + zq_0,$$

and $\bar{\eta}_{\mathcal{N}}(\alpha,\lambda_{\mathcal{N}})$ is the solution to

$$-\ln(1-\alpha) = \ln(1+\frac{z}{\lambda_{\mathcal{N}}}) + zq_0,$$

it is easy to verify that

$$\frac{\sum_{i\in\mathcal{N}}\bar{\eta}_i(\alpha,\lambda_i)}{\sum_{i\in\mathcal{N}}\lambda_i}\geq\min_{i\in\mathcal{N}}\frac{\bar{\eta}_i(\alpha,\lambda_i)}{\lambda_i}.$$

Hence, since $\bar{\eta}_i(\alpha,\lambda)/\lambda$ is nonincreasing in λ for all i, we must have

$$\sum_{i\in\mathcal{N}}\bar{\eta}_i(\alpha,\lambda_i)>\bar{\eta}_{\mathcal{N}}(\alpha,\lambda_{\mathcal{N}}).$$

To prove that $\sum_{i\in\mathcal{N}}\bar{\mu}_i^*\geq\bar{\mu}_{\mathcal{N}}^*$, note that

$$\begin{aligned}\sum_{i\in\mathcal{N}}\bar{\mu}_i^*-\lambda_{\mathcal{N}}&=\sum_{i\in\mathcal{N}}\max\{\bar{\eta}_i(\alpha,\lambda_i),\sqrt{\frac{h_i\lambda_i}{c}}\}\geq\max\{\sum_{i\in\mathcal{N}}\bar{\eta}_i(\alpha,\lambda_i),\sum_{i\in\mathcal{N}}\sqrt{\frac{h_i\lambda_i}{c}}\}\\&\geq\max\{\bar{\eta}_{\mathcal{N}}(\alpha,\lambda_{\mathcal{N}}),\sqrt{\frac{\sum_{i\in\mathcal{N}}h_i\lambda_i}{c}}\}=\bar{\mu}_{\mathcal{N}}^*-\lambda_{\mathcal{N}}.\end{aligned}$$

In order to prove that $\bar{z}_{\mathcal{N}}^*\leq\bar{z}_{1,\ldots,n}^*$, we distinguish two cases.

(1) $\bar{\eta}_{\mathcal{N}}(\alpha,\lambda_{\mathcal{N}})\leq\sqrt{\frac{\sum_{i\in\mathcal{N}}h_i\lambda_i}{c}}$. In this case, we have

$$\begin{aligned}\bar{z}_{\mathcal{N}}^*=&\frac{\sum_{i\in\mathcal{N}}h_i\lambda_i}{\sqrt{\frac{\sum_{i\in\mathcal{N}}h_i\lambda_i}{c}}}+c\sum_{i\in\mathcal{N}}\lambda_i+c\sqrt{\frac{\sum_{i\in\mathcal{N}}h_i\lambda_i}{c}}\leq\sum_{i\in\mathcal{N}}(\frac{h_i\lambda_i}{\sqrt{\frac{h_i\lambda_i}{c}}}+c\lambda_i+c\sqrt{\frac{h_i\lambda_i}{c}})\\\leq&\sum_{i\in\mathcal{N}}(c(\lambda_i+\bar{\eta}_i)+\frac{h_i\lambda_i}{\bar{\eta}_i})=\bar{z}_{1,\ldots,n}^*.\end{aligned}$$

(2) $\bar{\eta}_{\mathcal{N}}(\alpha,\lambda_{\mathcal{N}})>\sqrt{\frac{\sum_{i\in\mathcal{N}}h_i\lambda_i}{c}}$. In this case, we have

$$\begin{aligned}\bar{z}_{\mathcal{N}}^*=&\frac{\sum_{i\in\mathcal{N}}h_i\lambda_i}{\bar{\eta}_{\mathcal{N}}(\alpha,\lambda_{\mathcal{N}})}+c\sum_{i\in\mathcal{N}}\lambda_i+c\bar{\eta}_{\mathcal{N}}(\alpha,\lambda_{\mathcal{N}})\\\leq&\frac{\sum_{i\in\mathcal{N}}h_i\lambda_i}{\sum_{i\in\mathcal{N}}\bar{\eta}_i}+c\sum_{i\in\mathcal{N}}\lambda_i+c\sum_{i\in\mathcal{N}}\bar{\eta}_i\\\leq&\sum_{i\in\mathcal{N}}(c(\lambda_i+\bar{\eta}_i)+\frac{h_i\lambda_i}{\bar{\eta}_i})=\bar{z}_{1,\ldots,n}^*.\end{aligned}$$

where the first inequality is due to $\bar{\eta}_i=\max\{\eta_i(\alpha,\lambda_i),\sqrt{\frac{h_i\lambda_i}{c}}\}$ and $\sum_{i\in\mathcal{N}}\bar{\eta}_i\geq\eta_{\mathcal{N}}(\alpha,\lambda_{\mathcal{N}})>\sqrt{\frac{\sum_{i\in\mathcal{N}}h_i\lambda_i}{c}}$.

■

Proof of Theorem 4.5: We distinguish two cases.

(1) $\bar{\eta}_{\mathcal{N}}(\alpha,\lambda_{\mathcal{N}})\leq\sqrt{\frac{\sum_{i\in\mathcal{N}}h_i\lambda_i}{c}}$. First note that

$$z_{\mathcal{J}}^*\geq c\sum_{i\in\mathcal{J}}\lambda_i+2\sqrt{c\sum_{i\in\mathcal{J}}h_i\lambda_i}.$$

Since $\sum_{i\in\mathcal{J}}\phi_i - \left[c\sum_{i\in\mathcal{J}}\lambda_i + 2\sqrt{c\sum_{i\in\mathcal{J}}h_i\lambda_i}\right] = 2\left[\sum_{i\in\mathcal{J}}h_i\lambda_i\sqrt{\frac{c}{\sum_{i\in\mathcal{N}}h_i\lambda_i}} - \sqrt{c\sum_{i\in\mathcal{J}}h_i\lambda_i}\right] \leq 0$, we have $z^*_{\mathcal{J}} \geq \sum_{i\in\mathcal{J}}\phi_i$, $\forall\mathcal{J}\subseteq\mathcal{N}$. It follows that the allocation rule is in the core.

(2) $\bar{\eta}_{\mathcal{N}}(\alpha,\lambda_{\mathcal{N}}) > \sqrt{\frac{\sum_{i\in\mathcal{N}}h_i\lambda_i}{c}}$. Note that

$$\begin{aligned}&\frac{\sum_{i\in\mathcal{J}}h_i\lambda_i}{\bar{\eta}_{\mathcal{N}}(\alpha,\lambda_{\mathcal{N}})} + c\frac{\sum_{i\in\mathcal{J}}\lambda_i}{\lambda_{\mathcal{N}}}\bar{\eta}_{\mathcal{N}}(\alpha,\lambda_{\mathcal{N}})\\ \leq&\frac{\sum_{i\in\mathcal{J}}h_i\lambda_i}{\bar{\eta}_{\mathcal{J}}} + c\bar{\eta}_{\mathcal{J}}.\end{aligned}$$

The inequality is due to $\bar{\eta}_{\mathcal{N}}(\alpha,\lambda_{\mathcal{N}}) \geq \bar{\eta}_{\mathcal{J}}$ and $\frac{\bar{\eta}_{\mathcal{N}}(\alpha,\lambda_{\mathcal{N}})}{\lambda_{\mathcal{N}}} \leq \frac{\bar{\eta}_{\mathcal{J}}(\alpha,\lambda_{\mathcal{J}})}{\lambda_{\mathcal{J}}}$ by Lemma 3.1. Consequently, the allocation is in the core.

■

Proof of Lemma 5.1:

In order to utilize the achievable region method to analyze our problem, we first introduce the notion of *polymatroid*. Let $\mathcal{N}=\{1,\cdots,n\}$. A function $g: 2^{\mathcal{N}}\to\mathbb{R}$ is defined as a rank function if it has the following properties: (1) $g(\phi)=0$; (2) g is nondecreasing, i.e., $g(\mathcal{S})\leq g(\mathcal{T})$ for any $\mathcal{S}\subseteq\mathcal{T}\subseteq\mathcal{N}$; (3) g is supermodular, i.e., $g(\mathcal{J}\cup\mathcal{T})+g(\mathcal{J}\cap\mathcal{T})\geq g(\mathcal{J})+g(\mathcal{T})$. Given a subset $T\subseteq\mathcal{N}$ and a function $g: 2^{\mathcal{N}}\to\mathbb{R}$, the following polyhedron

$$\{(x_i, i\in\mathcal{T}) : \sum_{i\in\mathcal{V}} x_i \geq g(\mathcal{V}), \forall\mathcal{V}\subseteq\mathcal{T}; x_i\geq 0, i\in\mathcal{T}\}$$

is called a *polymatroid* if g is a rank function (see Edmonds 1970).

Define $\pi_i=\{1,\cdots,i\}$. Consider the following linear programming problem.

$$\begin{aligned}&\min\sum_{i=1}^{n} c_i x_i\\ \text{subject to }&\sum_{i\in\mathcal{V}} x_i \geq g(\mathcal{V}), \forall\mathcal{V}\subseteq\mathcal{N},\\ &x_i\geq 0, i\in\mathcal{N},\end{aligned} \tag{31}$$

where g is a rank function.

Lemma 7.1 (Edmonds 1970) *Suppose that $c_1\geq c_2\geq\cdots\geq c_n$. Then $x^*_1 = g(\pi_1), x^*_i = g(\pi_i) - g(\pi_{i-1}), i=2,\cdots,n$ is an optimal solution to (31). Moreover, if $c_1>c_2>\cdots>c_n$, then this is the unique optimal solution.*

Hence, if we optimize a linear objective function over a polymatriod, then a greedy algorithm is optimal.

Note that the feasible region of (22) is a polymatroid. Suppose that $(\xi_1,\cdots,\xi_{|\mathcal{J}|})$ is a permutation of the set of indices $\{i : i\in\mathcal{J}\}$ such that $h_{\xi_1}+\theta^*_{\xi_1,\mathcal{J}}\geq\cdots\geq h_{\xi_{|\mathcal{J}|}}+\theta^*_{\xi_{|\mathcal{J}|},\mathcal{J}}$. Let $\hat{\pi}_i=\{\xi_1,\cdots,\xi_i\}$. Based on

Lemma 7.1, $\lambda_{\xi_i} E[W^P_{\xi_i,\mathcal{J}}] = \frac{\lambda_{\hat{\pi}_i}}{\mu-\lambda_{\hat{\pi}_i}} - \frac{\lambda_{\hat{\pi}_{i-1}}}{\mu-\lambda_{\hat{\pi}_{i-1}}}, i = 2, \cdots, |\mathcal{J}|$ and $\lambda_{\xi_1} W^P_{\xi_1,\mathcal{J}} = \frac{\lambda_{\hat{\pi}_1}}{\mu-\lambda_{\hat{\pi}_1}}$. The optimal priority policy for coalition $\mathcal{J}$ is a strict preemptive priority policy and the priority order is decreasing in $h_i + \theta^*_{i,\mathcal{J}}$, with the firm with the highest value of $h_i + \theta^*_{i,\mathcal{J}}$ having the highest priority, which is due to that $\frac{\lambda_{\mathcal{V}}}{\mu-\lambda_{\mathcal{V}}}$ is nondecreasing and supermodular in $\mathcal{V}$. ■

Proof of Theorem 5.2:

Without loss of generality, in the following we assume that $\lambda_{\mathcal{N}} \leq \frac{1}{4}\frac{h_n}{c}$. This can be accomplished by rescaling the time (i.e., using a small enough time unit) since $\lambda_{\mathcal{N}}$ is the expected total arrival rate per unit of time and $\frac{h_n}{c}$ is unitless.

First, we consider the case without service level constraints. In this case the objective function is given by $h_1\frac{\lambda_{\pi_1\cap\mathcal{J}}}{\mu-\lambda_{\pi_1\cap\mathcal{J}}} + \sum_{i=2}^n h_i[\frac{\lambda_{\pi_i\cap\mathcal{J}}}{\mu-\lambda_{\pi_i\cap\mathcal{J}}} - \frac{\lambda_{\pi_{i-1}\cap\mathcal{J}}}{\mu-\lambda_{\pi_{i-1}\cap\mathcal{J}}}] + c\mu = \sum_{i=1}^n \gamma_i(\frac{\lambda_{\pi_i\cap\mathcal{J}}}{\mu-\lambda_{\pi_i\cap\mathcal{J}}}) + c\mu$, where $\gamma_i = h_i - h_{i+1} \geq 0, i = 1, \cdots, n-1$ and $\gamma_n = h_n$. As a result, the optimal capacity decision for coalition $\mathcal{J}$ is given by $\min_{\mu>\lambda_{\mathcal{J}}}[\sum_{i=1}^n \gamma_i(\frac{\lambda_{\pi_i\cap\mathcal{J}}}{\mu-\lambda_{\pi_i\cap\mathcal{J}}}) + c\mu]$.

Define

$$z^*(\lambda_1, \cdots, \lambda_n) = \min_{\mu>\lambda_{\mathcal{N}}}[\sum_{i=1}^n \gamma_i(\frac{\lambda_{\pi_i}}{\mu-\lambda_{\pi_i}}) + c\mu],$$

as the optimal cost for the grand coalition under demand rates $\lambda_1, \cdots, \lambda_n$. Next, we will show that z^* is submodular in $(\lambda_1, \cdots, \lambda_n)$. But the objective function inside the minimization operator is not submodular in $(\mu, \lambda_1, \cdots, \lambda_n)$. Instead we show it by using the envelope theorem. Let μ^* be the optimal capacity. First, it is clear that μ^* is nondecreasing in $h_1, \cdots, h_n$ since the objective function is submodular in $(h_1, \cdots, h_n, \mu)$. It follows that $\sqrt{\frac{h_n\lambda_{\mathcal{N}}}{c}} + \lambda_{\mathcal{N}} \leq \mu^* \leq \sqrt{\frac{h_1\lambda_{\mathcal{N}}}{c}} + \lambda_{\mathcal{N}}$ (noting that if the holding costs for all firms are h_n, then $\mu^* = \sqrt{\frac{h_n\lambda_{\mathcal{N}}}{c}} + \lambda_{\mathcal{N}}$. But as $h_1, \cdots, h_n$ increase, the optimal capacity also increases and hence the upper bound of μ^* is $\sqrt{\frac{h_1\lambda_{\mathcal{N}}}{c}} + \lambda_{\mathcal{N}}$). Without loss of generality, we assume that $m, k \in \mathcal{N}$ and $m < k$. By the envelope theorem (see Milgrom and Segal 2002 for more details), we have

$$\frac{\partial z^*(\lambda_1, \cdots, \lambda_n)}{\partial \lambda_k} = \sum_{i=k}^n \gamma_i(\frac{1}{\mu^* - \lambda_{\pi_i}}) + \sum_{i=k}^n \gamma_i(\frac{\lambda_{\pi_i}}{(\mu^* - \lambda_{\pi_i})^2}).$$

Then we have the following equation:

$$\begin{aligned}\frac{\partial^2 z^*(\lambda_1, \cdots, \lambda_n)}{\partial \lambda_k \partial \lambda_m} =& -\sum_{i=k}^n \gamma_i \frac{1}{(\mu^* - \lambda_{\pi_i})^2}(\frac{\partial \mu^*}{\partial \lambda_m} - 1) - \sum_{i=l}^n 2\gamma_i \frac{\lambda_{\pi_i}}{(\mu^* - \lambda_{\pi_i})^3}(\frac{\partial \mu^*}{\partial \lambda_m} - 1) \\ &+ \sum_{i=l}^n \gamma_i(\frac{1}{(\mu^* - \lambda_{\pi_i})^2}),\end{aligned}$$

since $m < k$. But $\frac{\partial \mu^*}{\partial \lambda_m} \geq \frac{\partial \sqrt{\frac{h_n\lambda_{\mathcal{N}}}{c}}}{\partial \lambda_m} + 1 = \frac{1}{2}\sqrt{\frac{h_n}{c\lambda_{\mathcal{N}}}} + 1$ based on the aforementioned properties of the optimal capacity (noting that $\sqrt{\frac{h_n\lambda_{\mathcal{N}}}{c}} + \lambda_{\mathcal{N}} \leq \mu^*$ and μ^* is nondecreasing in $h_1, \cdots, h_n$). Hence if $\frac{1}{2}\sqrt{\frac{h_n}{c\lambda_{\mathcal{N}}}} \geq 1$ or equivalently $\lambda_{\mathcal{N}} \leq \frac{1}{4}\frac{h_n}{c}$, then $\frac{\partial^2 z^*(\lambda_1, \cdots, \lambda_n)}{\partial \lambda_k \partial \lambda_m} \leq 0$ for all $k \neq m, k, m \in \mathcal{N}$. Therefore, the cooperative game

must be a submodular game.

Next, we show that the optimal cost is still submodular even with service level constraints. Suppose that $\mathcal{S}, \mathcal{T} \in \mathcal{N}$. Let $(\theta^*_{\mathcal{J}}, E[W^P_{i,\mathcal{J}}], i \in \mathcal{J}, \mu^P_{\mathcal{J}}), (\theta^*_{\mathcal{T}}, E[W^P_{i,\mathcal{T}}], i \in \mathcal{T}, \mu^P_{\mathcal{T}})$ be the optimal solutions for the subsets $\mathcal{J}$ and $\mathcal{T}$ for the Lagrangian problem in (23), respectively. Let $\hat{\theta}$ be the vector such that under $\hat{\theta}$, the optimal solution $E[\hat{W}^P_{i,\mathcal{J}\cup\mathcal{T}}], i \in \mathcal{J} \cup \mathcal{T}$ for

$$\begin{aligned} z(\hat{\theta}, \mathcal{J} \cup \mathcal{T}) = & \min_{E[W_i], i\in\mathcal{J}\cup\mathcal{T}, \mu} \sum_{i\in\mathcal{J}\cup\mathcal{T}} (h_i + \hat{\theta}_i)\lambda_i E[W_i] + c\mu - \sum_{i\in\mathcal{J}\cup\mathcal{T}} \hat{\theta}_i w_i \\ & \text{subject to } \sum_{i\in\mathcal{V}} \lambda_i E[W_i] \geq \frac{\lambda_{\mathcal{V}}}{\mu - \lambda_{\mathcal{V}}}, \mathcal{V} \subseteq \mathcal{J} \cup \mathcal{T}, \\ & E[W_i] \geq 0, i \in \mathcal{J} \cup \mathcal{T}, \\ & \mu > \lambda_{\mathcal{J}\cup\mathcal{T}}. \end{aligned}$$

satisfies $E[\hat{W}^P_{i,\mathcal{J}\cup\mathcal{T}}] \leq w_i, i \in \mathcal{J} \cup \mathcal{T}$ and the optimal solution $E[\hat{W}^P_{i,\mathcal{J}\cap\mathcal{T}}], i \in S \cap T$ for

$$\begin{aligned} z(\hat{\theta}, \mathcal{J} \cap \mathcal{T}) = & \min_{E[W_i], i\in\mathcal{J}\cap\mathcal{T}, \mu} \sum_{i\in\mathcal{J}\cap\mathcal{T}} (h_i + \hat{\theta}_i)\lambda_i E[W_i] + c\mu - \sum_{i\in\mathcal{J}\cap\mathcal{T}} \hat{\theta}_i w_i \\ & \text{subject to } \sum_{i\in\mathcal{V}} \lambda_i E[W_i] \geq \frac{\lambda_{\mathcal{V}}}{\mu - \lambda_{\mathcal{V}}}, \mathcal{V} \subseteq \mathcal{J} \cap \mathcal{T}, \\ & E[W_i] \geq 0, i \in \mathcal{J} \cap \mathcal{T}, \\ & \mu > \lambda_{\mathcal{J}\cap\mathcal{T}}. \end{aligned}$$

satisfies $E[\hat{W}^P_{i,\mathcal{J}\cap\mathcal{T}}] \leq w_i, i \in \mathcal{J} \cap \mathcal{T}$ (this is possible by setting $\hat{\theta}_i$ large enough and under which the optimal capacities are large enough since as $\min_i(h_i + \hat{\theta}_i) \to \infty$, the corresponding optimal capacity also approaches ∞).

Then it is clear that

$$\begin{aligned} & z^*_{\mathcal{J}} + z^*_{\mathcal{T}} \\ =& \sum_{i\in\mathcal{J}} (h_i + \theta^*_{i,\mathcal{J}})\lambda_i E[W^P_{i,\mathcal{J}}] + c\mu^P_{\mathcal{J}} - \sum_{i\in\mathcal{J}} \theta^*_{i,\mathcal{J}} w_i + \sum_{i\in\mathcal{T}} (h_i + \theta^*_{i,\mathcal{T}})\lambda_i E[W^P_{i,\mathcal{T}}] + c\mu^P_{\mathcal{T}} - \sum_{i\in\mathcal{T}} \theta^*_{i,\mathcal{T}} w_i \\ \geq& \sum_{i\in\mathcal{J}} (h_i + \hat{\theta}_i)\lambda_i E[W^P_{i,\mathcal{J}}] + c\mu^P_{\mathcal{J}} - \sum_{i\in\mathcal{J}} \hat{\theta}_i w_i + \sum_{i\in\mathcal{T}} (h_i + \hat{\theta}_i)\lambda_i E[W^P_{i,\mathcal{T}}] + c\mu^P_{\mathcal{T}} - \sum_{i\in\mathcal{T}} \hat{\theta}_i w_i \\ \geq& z(\hat{\theta}, \mathcal{J}) + z(\hat{\theta}, \mathcal{T}) \\ \geq& z(\hat{\theta}, \mathcal{J} \cup \mathcal{T}) + z(\hat{\theta}, \mathcal{J} \cap \mathcal{T}) \\ \geq& z^*_{\mathcal{J}\cup\mathcal{T}} + z^*_{\mathcal{J}\cap\mathcal{T}}. \end{aligned}$$

The first inequality is due to the fact that $\hat{\theta}$ is a feasible solution to either $\max_{\theta\geq 0}[\sum_{i\in\mathcal{J}}(h_i+\theta_i)\lambda_i E[W^P_{i,\mathcal{J}}] + c\mu^P_{\mathcal{J}} - \sum_{i\in\mathcal{J}} \theta_i w_i]$ or $\max_{\theta\geq 0}[\sum_{i\in\mathcal{T}}(h_i + \theta_i)\lambda_i E[W^P_{i,\mathcal{T}}] + c\mu^P_{\mathcal{T}} - \sum_{i\in T} \theta_i w_i]$; the second inequality is due to

the fact that $(E[W^P_{i,\mathcal{J}}], i \in \mathcal{J}, \mu^P_{\mathcal{J}})$ is a feasible solution to

$$\begin{aligned} z(\hat{\theta}, \mathcal{J}) = & \min_{E[W_i], i\in\mathcal{J};\mu} \sum_{i\in\mathcal{J}} (h_i + \hat{\theta}_i)\lambda_i E[W_i] + c\mu - \sum_{i\in\mathcal{J}} \hat{\theta}_i w_i \\ \text{subject to } & \sum_{i\in\mathcal{V}} \lambda_i E[W_i] \geq \frac{\lambda_{\mathcal{V}}}{\mu - \lambda_{\mathcal{V}}}, \mathcal{V} \subseteq \mathcal{J}, \\ & E[W_i] \geq 0, i \in \mathcal{J}, \end{aligned}$$

and $(E[W^P_{i,\mathcal{T}}], i \in \mathcal{T}, \mu^P_{\mathcal{T}})$ is a feasible solution to

$$\begin{aligned} z(\hat{\theta}, \mathcal{T}) = & \min_{E[W_i], i\in\mathcal{T},\mu} \sum_{i\in\mathcal{T}} (h_i + \hat{\theta}_i)\lambda_i E[W_i] + c\mu - \sum_{i\in\mathcal{T}} \hat{\theta}_i w_i \\ \text{subject to } & \sum_{i\in\mathcal{V}} \lambda_i E[W_i] \geq \frac{\lambda_{\mathcal{V}}}{\mu - \lambda_{\mathcal{V}}}, \mathcal{V} \subseteq \mathcal{T}, \\ & \lambda_i E[W_i] \geq 0, i \in \mathcal{T}; \end{aligned}$$

the third inequality is due to the fact that given any delay costs the optimal cost z is submodular (note that $\sum_{i\in S} \hat{\theta}_i w_i$ is a linear function of $\mathcal{J}$ and hence $z(\hat{\theta}, \mathcal{J})$ is submodular in $\mathcal{J}$ given $\hat{\theta} \geq 0$). Finally the fourth inequality is due to the fact that $z^*_{\mathcal{J}\cup\mathcal{T}}$ and $z^*_{\mathcal{J}\cap\mathcal{T}}$ are the optimal costs and under $\hat{\theta}$, the optimal solutions for coalitions $\mathcal{J} \cup \mathcal{T}$ and $\mathcal{J} \cap \mathcal{T}$ are feasible, and hence $z^*_{\mathcal{J}\cup\mathcal{T}} \leq z(\hat{\theta}, \mathcal{J} \cup \mathcal{T}), z^*_{\mathcal{J}\cap\mathcal{T}} \leq z(\hat{\theta}, \mathcal{J} \cap \mathcal{T})$.

Now we extend our results to systems with the non-preemptive priority policy. Under the optimal non-preemptive priority policy, the optimal cost is given by

$$\begin{aligned} \tilde{z}^*(\lambda_1, \cdots, \lambda_n) &= \min_{\mu>\lambda_{\mathcal{N}}} [\sum_{i=1}^n \gamma_i (\frac{\lambda_{\mathcal{N}}}{\mu} \frac{\lambda_{\pi_i}}{\mu - \lambda_{\pi_i}} + \frac{\lambda_{\pi_i}}{\mu}) + c\mu] \\ &= \min_{\mu>\lambda_{\mathcal{N}}} [\sum_{i=1}^n \gamma_i (\frac{\lambda_{\pi_i}}{\mu - \lambda_{\pi_i}} + \frac{\lambda_{\pi_i}\lambda_{\mathcal{N}\setminus\pi_i}}{\mu(\mu - \lambda_{\pi_i})}) + c\mu]. \end{aligned}$$

Let $\tilde{\mu}^*$ be the optimal capacity. Without loss of generality, we assume that $m, k \in \mathcal{N}$ and $m < k$. By the envelope theorem, we have

$$\begin{aligned} \frac{\partial \tilde{z}^*(\lambda_1, \cdots, \lambda_n)}{\partial \lambda_k} = & \sum_{i=k}^n \gamma_i (\frac{1}{\tilde{\mu}^* - \lambda_{\pi_i}} + \frac{\lambda_{\pi_i}}{(\tilde{\mu}^* - \lambda_{\pi_i})^2}) \\ & + \sum_{i=k}^n \gamma_i (\frac{\lambda_{\mathcal{N}\setminus\pi_i}}{\tilde{\mu}^*(\tilde{\mu}^* - \lambda_{\pi_i})} + \frac{\lambda_{\pi_i}\lambda_{\mathcal{N}\setminus\pi_i}}{\tilde{\mu}^*(\tilde{\mu}^* - \lambda_{\pi_i})^2}). \end{aligned}$$

Based on the foregoing result,

$$\begin{aligned}
&\frac{\partial \tilde{z}^*(\lambda_1,\cdots,\lambda_n)}{\partial\lambda_k\partial\lambda_m}\\
=&\sum_{i=k}^{n}\gamma_i(-\frac{1}{(\tilde{\mu}^*-\lambda_{\pi_i})^2}(\frac{\partial\tilde{\mu}^*}{\partial\lambda_m}-1)+\frac{1}{(\tilde{\mu}^*-\lambda_{\pi_i})^2}-2\frac{\lambda_{\pi_i}}{(\tilde{\mu}^*-\lambda_{\pi_i})^3}(\frac{\partial\tilde{\mu}^*}{\partial\lambda_m}-1))\\
&+\sum_{i=k}^{n}\gamma_i(-\frac{\lambda_{\mathcal{N}\setminus\pi_i}}{(\tilde{\mu}^*)^2(\tilde{\mu}^*-\lambda_{\pi_i})}\frac{\partial\tilde{\mu}^*}{\partial\lambda_m}-\frac{\lambda_{\mathcal{N}\setminus\pi_i}}{(\tilde{\mu}^*)(\tilde{\mu}^*-\lambda_{\pi_i})^2}(\frac{\partial\tilde{\mu}^*}{\partial\lambda_m}-1)+\frac{\lambda_{\mathcal{N}\setminus\pi_i}}{\tilde{\mu}^*(\tilde{\mu}^*-\lambda_{\pi_i})^2}\\
&-\frac{\lambda_{\pi_i}\lambda_{\mathcal{N}\setminus\pi_i}}{(\tilde{\mu}^*)^2(\tilde{\mu}^*-\lambda_{\pi_i})^2}\frac{\partial\tilde{\mu}^*}{\partial\lambda_m}-2\frac{\lambda_{\pi_i}\lambda_{\mathcal{N}\setminus\pi_i}}{(\tilde{\mu}^*)(\tilde{\mu}^*-\lambda_{\pi_i})^3}(\frac{\partial\tilde{\mu}^*}{\partial\lambda_m}-1)),
\end{aligned}$$

since $m<k$. It is clear that $\sqrt{\frac{h_n\lambda_{\mathcal{N}}}{c}}+\lambda_{\mathcal{N}}\le\tilde{\mu}^*\le\sqrt{\frac{h_1\lambda_{\mathcal{N}}}{c}}+\lambda_{\mathcal{N}}$ and $\frac{\partial\tilde{\mu}^*}{\partial\lambda_m}\ge\frac{\partial\sqrt{\frac{h_n\lambda_{\mathcal{N}}}{c}}}{\partial\lambda_m}+1=\frac{1}{2}\sqrt{\frac{h_n}{c\lambda_{\mathcal{N}}}}+1$ based on a similar argument on the proof for systems under preemptive priority policies. Hence if $\frac{1}{2}\sqrt{\frac{h_n}{c\lambda_{\mathcal{N}}}}\ge 1$ or equivalently $\lambda_{\mathcal{N}}\le\frac{1}{4}\frac{h_n}{c}$, then $\frac{\partial^2\tilde{z}^*(\lambda_1,\cdots,\lambda_n)}{\partial\lambda_k\partial\lambda_m}\le 0$ for all $k\neq m,k,m\in\mathcal{N}$. Hence, the corresponding cooperative game must be submodular. Based on Shanthikumar and Yao (1992), i.e., the performance measures for an M/M/1 system under the nonpreemptive priority policy can still have an achievable region. Then the analysis for the case with service level constraints on the expected delay can be accomplished by a similar argument as in the case for the systems with the preemptive priority policy. ■

Proof of Theorem 6.1:

First, we can show that $(\frac{h_i}{l_i}-\frac{h_j}{l_j})(l_i-l_j)\ge 0,\forall i\neq j$, which is equivalent to $(h_il_j-h_jl_i)(l_i-l_j)\le 0,\forall i\neq j$, imply that

$$\frac{\sum_{i\in\mathcal{N}}h_i\lambda_i}{\sum_{i\in\mathcal{N}}h_i\lambda_il_i}\le\frac{\sum_{i\in\mathcal{N}}\lambda_il_i}{\sum_{i\in\mathcal{N}}\lambda_il_i^2}. \tag{32}$$

Then it is sufficient to show that the total expected delay cost in the independent system, given capacities $\mu_i>\lambda_i$ for $i\in\mathcal{N}$, is greater than or equal to the expected delay cost in a shared system with capacity $\mu=\sum_{i\in\mathcal{N}}\mu_i$, i.e.,

$$\sum_{i\in\mathcal{N}}\frac{h_i\lambda_il_i}{\mu_i-\lambda_il_i}\ge\frac{(\sum_{i\in\mathcal{N}}h_i\lambda_i)(\sum_{i\in\mathcal{N}}\lambda_il_i^2)}{\mu(\mu-\sum_{i\in\mathcal{N}}\lambda_il_i)}+\frac{\sum_{i\in\mathcal{N}}h_i\lambda_il_i}{\mu}.$$

Then note that with (32) we have

$$\begin{aligned}
&\frac{(\sum_{i\in\mathcal{N}}h_i\lambda_i)(\sum_{i\in\mathcal{N}}\lambda_il_i^2)}{\mu(\mu-\sum_{i\in\mathcal{N}}\lambda_il_i)}+\frac{\sum_{i\in\mathcal{N}}h_i\lambda_il_i}{\mu}\\
=&\frac{\mu\sum_{i\in\mathcal{N}}h_i\lambda_il_i+\sum_{i\in\mathcal{N}}h_i\lambda_i\sum_{i\in\mathcal{N}}\lambda_il_i^2-\sum_{i\in\mathcal{N}}h_i\lambda_il_i\sum_{i\in\mathcal{N}}\lambda_il_i}{\mu(\mu-\sum_{i\in\mathcal{N}}\lambda_il_i)}\\
\le&\frac{\sum_{i\in\mathcal{N}}h_i\lambda_il_i}{\mu-\sum_{i\in\mathcal{N}}\lambda_il_i}.
\end{aligned}$$

However, we have

$$\frac{\sum_{i\in\mathcal{N}}h_i\lambda_il_i}{\sum_{i\in\mathcal{N}}\mu_i-\sum_{i\in\mathcal{N}}\lambda_il_i}\le\max_i\{\frac{h_i\lambda_il_i}{\mu_i-\lambda_il_i}\}\le\sum_{i\in\mathcal{N}}\frac{h_i\lambda_il_i}{\mu_i-\lambda_il_i},$$

which completes the proof.

If $h_i/h_j = l_i/l_j$ for all i, j, then $\frac{(\sum_{i\in\mathcal{N}} h_i\lambda_i)(\sum_{i\in\mathcal{N}} \lambda_i l_i^2)}{\mu(\mu-\sum_{i\in\mathcal{N}} \lambda_i l_i)} + \frac{\sum_{i\in\mathcal{N}} h_i\lambda_i l_i}{\mu} = \frac{\sum_{i=1}^n h_i\lambda_i l_i}{\mu-\sum_{i\in\mathcal{N}} \lambda_i l_i}$. Arguments similar to those made in the proof of the M/M/1 case complete the proof. ■

Proof to Theorem 6.3: A sufficient condition for (29) to hold is for

$$\sqrt{\frac{\sum_{i=1}^n \lambda_i\sigma_i^2}{\sum_{i=1}^n \lambda_i}}\sqrt{\sum_{i=1}^n h_i\lambda_i} \le \sum_{i=1}^n \sqrt{h_i\lambda_i\sigma_i^2},$$

or equivalently, by taking the square of both sides of the inequality, for

$$\frac{\sum_{i=1}^n \lambda_i\sigma_i^2 \sum_{i=1}^n h_i\lambda_i}{\sum_{i=1}^n \lambda_i} \le \sum_{i=1}^n h_i\lambda_i\sigma_i^2 + 2\sum_{j\neq i}\sqrt{h_i\lambda_i\sigma_i^2 h_j\lambda_j\sigma_j^2}.$$

To prove that $\sum_{i=1}^n \lambda_i \left(\sum_{i=1}^n h_i\lambda_i\sigma_i^2 + 2\sum_{j\neq i}\sqrt{h_i\lambda_i\sigma_i^2 h_j\lambda_j\sigma_j^2}\right) \ge \sum_{i=1}^n \lambda_i\sigma_i^2 \sum_{i=1}^n h_i\lambda_i$, it is sufficient to show that $\sum_{i=1}^n \lambda_i \sum_{i=1}^n h_i\lambda_i\sigma_i^2 \ge \sum_{i=1}^n \lambda_i\sigma_i^2 \sum_{i=1}^n h_i\lambda_i$. Note that

$$\sum_{i=1}^n \lambda_i \sum_{i=1}^n h_i\lambda_i\sigma_i^2 = \sum_{i=1}^n h_i\lambda_i^2\sigma_i^2 + 2\sum_{j\neq i} h_i\lambda_i\lambda_j\sigma_i^2$$

and

$$\sum_{i=1}^n \lambda_i\sigma_i^2 \sum_{i=1}^n h_i\lambda_i = \sum_{i=1}^n h_i\lambda_i^2\sigma_i^2 + 2\sum_{j\neq i} h_i\lambda_i\lambda_j\sigma_j^2.$$

Now let $a(i,j) = h_i\lambda_i\sigma_i^2\lambda_j, a(j,i) = h_j\lambda_j\sigma_j^2\lambda_i$ and $b(i,j) = h_i\lambda_i\lambda_j\sigma_j^2, b(j,i) = h_j\lambda_j\lambda_i\sigma_i^2$. Then,

$$a(i,j) + a(j,i) - b(i,j) - b(j,i) = (h_i - h_j)\lambda_i\lambda_j(\sigma_i^2 - \sigma_j^2).$$

Hence if we have $(h_i - h_j)(\sigma_i^2 - \sigma_j^2) \ge 0$, we must also have

$$\sum_{j\neq i}(h_i - h_j)(\sigma_i^2 - \sigma_j^2) = \sum_{j\neq i} h_i\lambda_i\lambda_j\sigma_i^2 \ge \sum_{j\neq i} h_i\lambda_i\lambda_j\sigma_j^2 \ge 0.$$

Consequently, we also have $\sum_{i=1}^n \lambda_i \sum_{i=1}^n h_i\lambda_i\sigma_i^2 \ge \sum_{i=1}^n \lambda_i\sigma_i^2 \sum_{i=1}^n h_i\lambda_i$, which completes the proof.

Note that

$$\begin{aligned}
&\frac{\sum_{i\in\mathcal{J}} h_i\lambda_i}{\sum_{i=1}^n h_i\lambda_i}\sigma_{\mathcal{N}}\sqrt{\frac{\sum_{i=1}^n h_i\lambda_i}{c}} \\
=&\frac{\sum_{i\in\mathcal{J}} h_i\lambda_i}{\sum_{i=1}^n h_i\lambda_i}\sqrt{\frac{\sum_{i=1}^n \lambda_i\sigma_i^2}{\sum_{i=1}^n \lambda_i}}\sqrt{\frac{\sum_{i=1}^n h_i\lambda_i}{c}} \\
\le&\sqrt{\frac{\sum_{i\in S} \lambda_i\sigma_i^2}{\sum_{i\in\mathcal{J}} \lambda_i}}\sqrt{\frac{\sum_{i\in\mathcal{J}} h_i\lambda_i}{c}},
\end{aligned}$$

where the inequality is due to the assumption in the theorem. Hence,

$$\begin{aligned}
\sum_{i\in\mathcal{J}}\phi_i =& \sum_{i\in\mathcal{J}}\lambda_i + \sigma_{\mathcal{N}}\frac{\sum_{i\in\mathcal{J}} h_i\lambda_i}{\sqrt{\frac{\sum_{i=1}^n h_i\lambda_i}{c}}} + c\frac{\sum_{i\in\mathcal{J}} h_i\lambda_i}{\sum_{i=1}^n h_i\lambda_i}\sigma_{\mathcal{N}}\sqrt{\frac{\sum_{i=1}^n h_i\lambda_i}{c}} \\
=& \sum_{i\in\mathcal{J}}\lambda_i + 2c\frac{\sum_{i\in\mathcal{J}} h_i\lambda_i}{\sum_{i=1}^n h_i\lambda_i}\sigma_{\mathcal{N}}\sqrt{\frac{\sum_{i=1}^n h_i\lambda_i}{c}} \\
\leq& \sum_{i\in\mathcal{J}}\lambda_i + 2c\sqrt{\frac{\sum_{i\in\mathcal{J}}\lambda_i\sigma_i^2}{\sum_{i\in\mathcal{J}}\lambda_i}}\sqrt{\frac{\sum_{i\in\mathcal{J}} h_i\lambda_i}{c}} \\
=& z^*_{\mathcal{J}}.
\end{aligned}$$

■

Appendix B

The Benefits of Capacity Sharing

In this section, we numerically investigate the benefits of capacity sharing under both the FCFS policy and the optimal preemptive priority policy.

First, we consider the case with 3 firms. There are 3 systems: the independent system; the shared system under the FCFS policy and the shared system under the optimal preemptive priority policy. Specifically, the arrival rates are $\lambda_1 = 10\gamma, \lambda_2 = 20\gamma, \lambda_3 = 30\gamma$, and the delay costs are given by $\beta, 2\beta, 3\beta$, respectively. All firms subject to minimum expected delays $0.2\theta, 0.2\theta, 0.3\theta$, respectively. In the following three figures, the vertical axis denotes the percentage of cost saving of a shared system with respect to that of the independent system.

In Figure 1, we can see that as γ increases (hence all demand rates increase), then benefits of capacity sharing decrease for both the FCFS system and the priority system. This implies that capacity sharing is more effective when demand rates are small. Moreover, the benefits in are more significant when the minimum delay requirements are smaller. Hence, firms are more willing to participate capacity sharing when they have very stringent service level constraints, which is reasonable since capacity sharing can take advantage of the risk pooling.

In Figure 2, as β increases (hence all delay costs increase), then the benefits of capacity sharing mostly have an upward trend, i.e., as the delay costs increase, the benefits of capacity sharing also increase. Moreover, the benefits of capacity sharing are more significant when γ is smaller, which is consistent with Figure 1. But in priority system 2, as we can see that the benefit of capacity sharing is not monotone with respect to β. Hence, how the delay costs affect the benefit of capacity sharing should be taken with caution.

In Figure 3, as the capacity cost c increases, the benefits of capacity sharing decrease and when c is large enough these benefits converge to a constant level. This is due to that as capacity cost is large enough, then the shared system cannot take full advantage of the risk pooling effect. In addition, mostly the benefits of capacity sharing are more significant when β is larger.

Finally, in each of all these three figures, the benefit of capacity sharing for the priority system is more significant than that of the FCFS system. But the gap is not as big as the benefit of capacity sharing for the FCFS system. In summary, in general firms have more incentive to participate in capacity sharing when the delay costs are large, demand rates and capacity costs are not so large and service level constraints are very stringent.

Numerical Examples for Capacity Sharing with Heterogeneous Work Contents

It is difficult to characterize analytically the optimal capacity in a shared facility. Therefore, it is also difficult to characterize analytically conditions under which capacity sharing is beneficial (and the core may or may

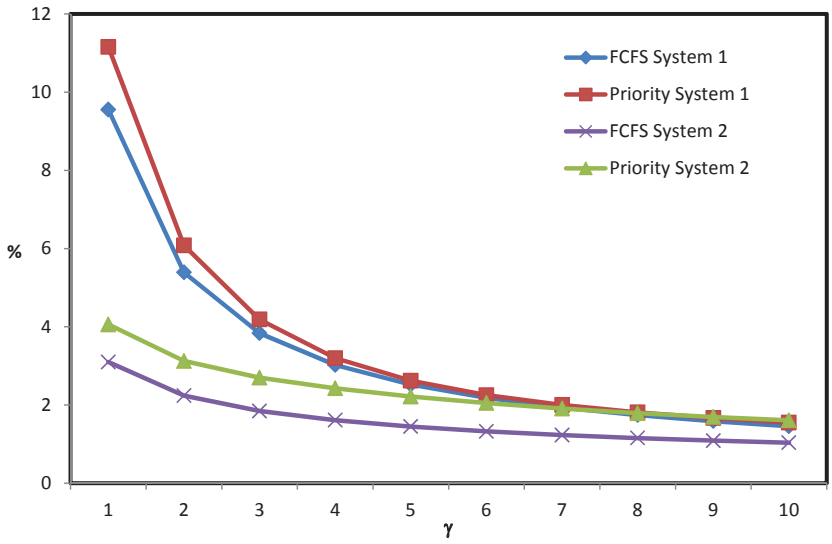

Figure 1: The Benefits of Capacity Sharing ($\beta = 1, c = 50$; System 1: $\theta = 1$; System 2: $\theta = 10$)

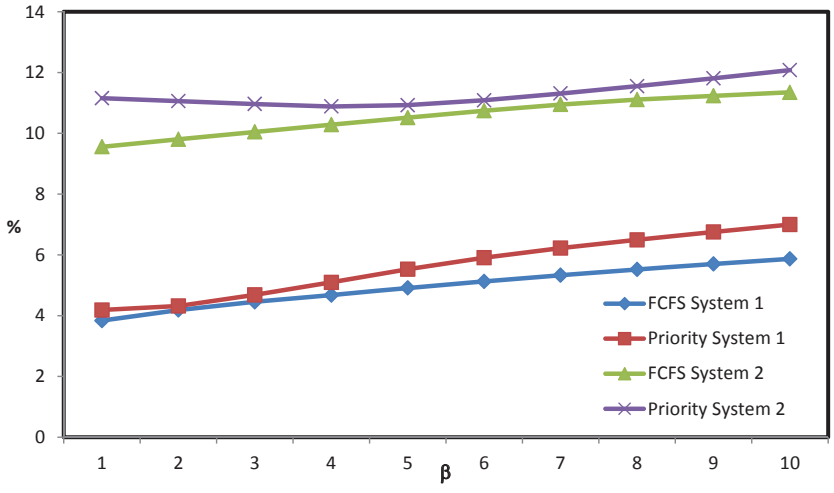

Figure 2: The Benefits of Capacity Sharing ($\theta = 1, c = 50$; System 1: $\gamma = 1$; System 2: $\gamma = 10$)

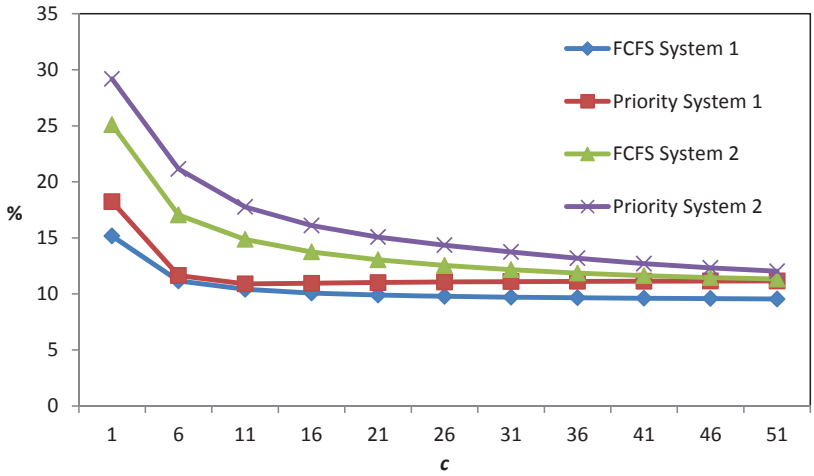

Figure 3: The Benefits of Capacity Sharing ($\gamma = 1, \theta = 1$; System 1: $\beta = 1$; System 2: $\beta = 10$)

not be empty). In order to get some insights, and isolate the effect of work content variability, let us consider the case where $h_i = h$ and $\lambda_i = \lambda$ for all $i \in \mathcal{N}$. The expression for C_s^2 then simplifies to $C_s^2 = 2C_l^2 + 1$ where $C_l = \sqrt{\frac{\sum_{i\in\mathcal{N}} nl_i^2 - (\sum_{i\in\mathcal{N}} l_i)^2}{(\sum_{i\in\mathcal{N}} l_i)^2}}$ is the coefficient of variation of the means of the individual work contents. The total cost in the shared system can now be expressed as

$$z_{\mathcal{N}}(\mu_{\mathcal{N}}) = c\mu_{\mathcal{N}} + h\left[\frac{(1+C_l^2)\rho^2}{(1-\rho)} + \rho\right].$$

It is easy to see that, everything else remaining the same, the total cost for a given service rate $\mu_{\mathcal{N}}$ is increasing in C_l and, consequently, the optimal capacity level is also increasing in C_l. This suggests that for sufficiently high C_l, the shared system could be less desirable than a system with independent facilities. This observation is confirmed by the numerical results shown in Figure 1 where the percentage cost difference between systems with and without capacity sharing, $(z^*_{\mathcal{N}} - z^*_{1,\ldots,n})/z^*_{1,\ldots,n}$, is shown for different values of C_l for an example system with two firms (to obtain different values of C_l, we vary l_1 and l_2 while keeping $l_1 + l_2 = 20$). There is a threshold on C_l above which the system without capacity sharing becomes preferable. The effect of increasing C_l is particularly significant when delay cost is high (as shown in Figure 1) or when the demand rates or capacity cost are low (as observed in additional numerical results not shown here for brevity). these are the settings where the difference between pooled and distributed systems tends to be most significant.

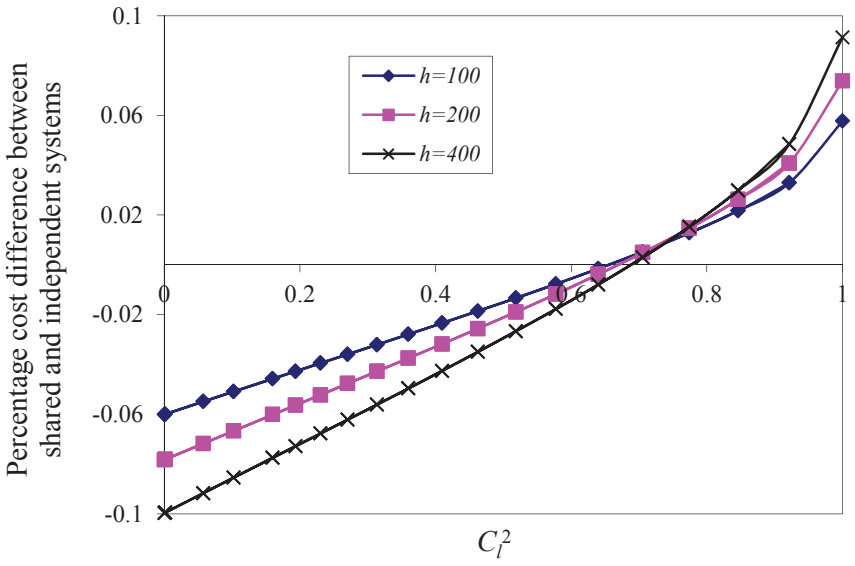

Figure 4 – The effect of work content variability
($c = 30, \lambda_1 = \lambda_2 = 20, l_1 + l_2 = 20$)

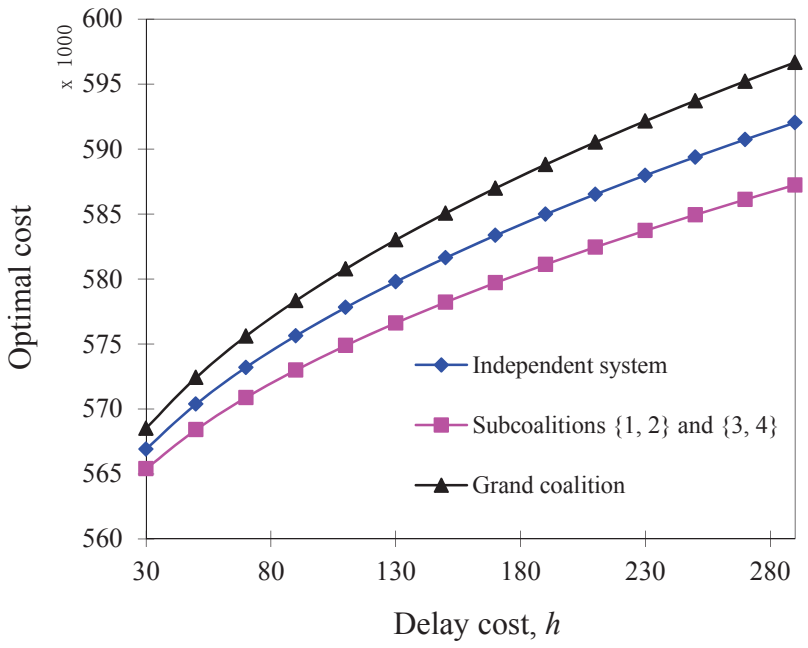

Figure 5 – The benefit of partial pooling
($c = 50, \lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 20, l_1 = l_2 = 2.5, l_3 = 100, l_4 = 450$)